DOCUMENT RESUME

ED 470 225   TM 034 573

AUTHOR   Glass, Gene V., Ed.
TITLE   Education Policy Analysis Archives, 2002: Numbers 1-25.
ISSN   ISSN-1068-2341
PUB DATE   2002-00-00
NOTE   392p.; "Education Policy Analysis Archives" is an electronic-
               only journal (see http://olam.edu.edu/epaa) covered on an
               article-by-article basis in "Current Index to Journals in
               Education" (CIJE). For the 24 English-language articles in
               this part of Volume 10, see TM 524 637-660 (in CIJE).
PUB TYPE   Collected Works - Serials (022)
JOURNAL CIT   Education Policy Analysis Archives; 10 n1-25 2002
EDRS PRICE   EDRS Price MF01/PC16 Plus Postage.
DESCRIPTORS   *Academic Achievement; *Educational Change; *Elementary
               Secondary Education; Equal Education; High Stakes Tests;
               *Higher Education; State Programs; *Test Use; *Testing
               Programs

ABSTRACT

          This document consists of articles 1 through 25 published in
the electronic journal "Education Policy Analysis Archives" for the year
2002: (1) "Testing and Diversity in Postsecondary Education: The Case of
California" (Daniel Koretz, Michael Russell, Chingwei David Shin, Cathy Horn,
and Kelly Shasby); (2) "State-Mandated Testing and Teachers' Beliefs and
Practice" (Sandra Cimbricz); (3) "Socratic Pedagogy, Race, and Power: From
People to Propositions" (Peter Boghossian); (4) "Technology Is Changing
What's 'Fair Use' in Teaching—Again" (Linda Howe-Stegier and Brian C.
Donohue); (5) "The Power-discourse Relationship in a Croatian Higher
Education Setting" (Renata Fox and John Fox); (6) "Technical and Ethical
Issues in Indicator Systems: Doing Things Right and Doing Wrong
Things" (Carol Taylor Fitz-Gibbon and Peter Tymms); (7) "'Exito in
California?' A Validity Critique of Language Program Evaluations and Analysis
of English Learner Test Scores" (Marilyn S. Thompson, Kristen E. DiCerbo,
Kate Mahoney, and Jeff MacSwan); (8) "Reaction to Bolton's 'Significance of
Test-based Ratings for Metropolitan Boston Schools'" (Stephan Michelson); (9)
"Confusing the Messenger with the Message: A Response to Bolon" (Victor L.
Willson and Thomas Kellow); (10) "Response to Michelson and to Willson and
Kellow" (Craig Bolon); (11) "Women in Managerial Positions in Greek
Education: Evidence of Inequality" (Anastasia Athanassoula-Reppa and Manolis
Koutouzis); (12) "How Schools Matter: The Link between Teacher Classroom
Practices and Student Academic Performance" (Harold Wenglinsky); (13) "Los
Rituales Escolares y las Practicas Educativas" (Pablo Daniel Vain); (14)
"Japanese Higher Education Policy in Korea during the Colonial Period (1919-
1945)" (Jeong-Kyu Lee); (15) "Basic Education Reform in China: Untangling the
Story of Success" (Chengzhi Wang and Quanhua Zhou); (16) "Quantifying
Quality: What Can the 'U.S. News and World Report' Rankings Tell Us about the
Quality of Higher Education" (Marguerite Clarke); (17) "Are Increasing Test
Scores in Texas Really a Myth, Or Is Haney's Myth a Myth?" (Laurence A.
Toenjes and A. Gary Dworkin); (18) "High-Stakes Testing, Uncertainty, and
Student Learning" (Audrey L. Amrein and David C. Berliner); (19)
"Entrepreneurial Ambitions in the Public Sector:A Random Effects Model of the
Emergence of Charter Schools in North Carolina" (Linda A. Renzulli); (20)
"The Amalgamation of Chinese Higher Education Institutions" (David Y. Chen);
(21) "The Possibility of Reform: Micropolitics in Higher Education" (Susan

Education Policy Analysis Archives

(Articles 1 thru 25)

editor:
Gene V. Glass
Arizona State University

TM034573

Education Policy Analysis Archives

How to Subscribe to the EPAA Listserv
EPAA is Distributed Free of All Charges

EPAA operates a LISTSERV at asu.edu at Arizona State University. **The purpose of this List is not to distribute articles that are published in the journal, but rather to announce the availability of new articles on the EPAA World Wide Web site.**

To subscribe, send an email letter to LISTSERV@asu.edu and make the one line contents read SUB EPAA "your name" (no quotes).

You may cancel your subscription to EPAA at any time by sending an e-mail letter to LISTSERV@asu.edu and including in the letter the single line SIGNOFF EPAA.

Cómo suscribirse a la lista *EPAA* Listserv
EPAA se distribuye en forma gratuita

EPAA opera una LISTSERV en asu.edu en la Arizona State University. El propósito de esta lista no es distribuir los artículos publicados en la revista, sino solamente anunciar la disponibilidad de artículos nuevos en el sitio WWW de *EPAA*.

Para suscribirse, envíe un mensaje email a LISTSERV@asu.edu y escriba en la primera línea el texto: SUB EPAA "su nombre" (sin comillas).

De preferencia envíe el mensaje email sin "signature"

Puede cancelar su subscripción a *EPAA* en cualquier momento al enviar un mensaje email a LISTSERV@asu.edu con la leyenda SIGNOFF EPAA

BEST COPY AVAILABLE

Education Policy Analysis Archives

EPAA Editors

EPAA is Edited by **Gene V Glass**, who is a Professor in the College of Education at Arizona State University. You may reach the Editor in a variety of ways:

- Email to glass@asu.edu
- Paper mail to College of Education, Arizona State University, Tempe AZ 85287-2411
- Voice messages are taken at 602-965-2692

The *EPAA* **Associate Editor for Spanish Language** is **Roberto Rodríguez Gómez** of the Universidad Nacional Autónoma de México. His email address is roberto@servidor.unam.mx. Articles in Spanish should be submitted to Associate Editor Rodríguez, who will supervise the review and make final editorial decisions on Spanish language articles.

The **Book Review Editor** for *EPAA* is Walter E. Shepherd. He can be reached at shepherd@asu.edu

The **Commentary Editor** for *EPAA* is Casey D. Cobb, an Assistant Professor in the Department of Education at the University of New Hampshire. He can be reached at casey.cobb@unh.edu

Articles published in *EPAA* are peer reviewed by members of the Editorial Board. Neither the Editors, their institutions, nor the Editorial Board members are responsible for errors of fact or interpretation that are not normally discoverable through the traditional process of scholarly peer review.

Education Policy Analysis Archives

EPAA English Language Editorial Board

BEST COPY AVAILABLE

6

Simon Schwartzman (Brazil)
American Institutes for Research

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles

7

Education Policy Analysis Archives

9

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Testing and Diversity in Postsecondary Education:
The Case of California

Daniel Koretz
Harvard University

Michael Russell
Boston College

Chingwei David Shin
University of Iowa

Cathy Horn
Harvard University

Kelly Shasby
Boston College

Abstract
The past several years have seen numerous efforts to scale back or eliminate affirmative action in
postsecondary admissions. In response, policymakers and postsecondary institutions in many
states are searching for ways to maintain the diversity of student populations without resorting to
a prohibited focus on race. In response to these changes, this study used data from California and
a simplified model of the University of California admissions process to explore how various
approaches to admissions affect the diversity of the admitted student population. "Race-neutral"
admissions based solely on test scores and grades were compared with the results of actual
admissions before and after the elimination of affirmative action. A final set of analyses explored
the effects on diversity of alternative approaches that take into account factors other than grades
and scores, but not race or ethnicity. Replacing the former admissions process that included
preferences with a race-neutral model based solely on GPA and SAT-I scores substantially
reduced minority representation at the two most selective UC campuses but had much smaller
effects at the other six, less selective campuses. SAT-I scores contributed to but were not the sole
cause of the underrepresentation of African American and Hispanic students. A race-neutral
model based solely on GPA also produced an underrepresentation of minorities, albeit a less
severe one. None of the alternative admissions models analyzed could replicate the composition
of the student population that was in place before the termination of affirmative action in
California. The only approach that substantially increased the representation of minority students

was accepting most students on the basis of within-school rather than statewide rankings, and this approach caused a sizable drop in both the average SAT scores and the average GPA of admitted applicants, particularly among African American and Hispanic students. Although admissions systems differ, the basic findings of this study are likely to apply at a general level to many universities and underscore the difficulty of providing proportional representation for underserved minority students at highly selective institutions without explicit preferences.

Over the past several years, the use of affirmative action to increase the representation of underserved minorities in postsecondary education has faced increasingly widespread threats. Efforts to scale back or eliminate affirmative action in admissions have taken numerous forms, including popular referenda, court decisions, and executive actions, and have led to its elimination in California, Florida, and Washington. Affirmative action was also terminated in Texas, although a recent court decision may permit its reinstatement. A university affirmative action program recently survived legal challenge in Michigan, but additional litigation is pending in other states, including cases on appeal in Michigan and Georgia.

Faced with these events, policymakers and postsecondary institutions in many states are searching for ways to maintain the diversity of student populations without resorting to a prohibited focus on race. For example, several states are experimenting with what we call "X% rules," in which the best students in each high school—that is, those who exceed a specified percentile in class rank—are guaranteed admission to a college campus or system.

In response to these changes, we undertook to explore how various approaches to admissions affect the diversity of the admitted student population. We examined the effects of different stages in the admissions process—for example, the student's decision to take a college-admissions test and to apply, and the college's decision to accept an applicant—on the composition of the student population. In doing this, we focused not only on race and ethnicity, but also on other aspects of diversity, such as the educational background of accepted students. We modeled "race-neutral" admissions based solely on test scores and grades and compared the results with actual admissions before and after the elimination of affirmative action. Finally, we explored the effects on diversity of alternative approaches that take into account factors other than grades and scores, but not race or ethnicity.

Because California has been one of the primary focuses of debate about the rollback of affirmative action, these analyses use data from that state and are loosely modeled after the admissions procedures and student population of the University of California. We expect that many of the findings could be generalized in broad brush to other states, but some patterns may differ depending on differences among states in demographics, the selectivity of state universities, and so on.

Recent Trends in Postsecondary Admissions

Although hard data on affirmative action are scanty, most observers believe that selective institutions have widely employed it for several decades. The use of race as a factor in higher education admissions was legitimized but limited by a 1978 Supreme Court decision, *University of California Board of Regents v. Bakke* (1978). The justices held that racial diversity was a legitimate goal for institutions of higher education but that creating a separate admissions process or a quota system was not the "least objectionable alternative" for achieving that goal. Consideration of race in the admissions process was deemed acceptable only if it was one of many factors considered.

The policies endorsing affirmative action have sometimes been explicit. An example is the state system of higher education in Texas. Until 1996, the state of Texas maintained a concerted effort to recruit minority students into higher education and to prepare its institutions to meet the demands of its growing minority college-age population. This effort was due to an investigation of segregation in Texas higher education by the U.S. Office of Civil Rights (OCR) between 1978 and 1981. As a result of that investigation, Texas was required to develop a plan to desegregate and to "increase the representation of blacks and Hispanics in institutions of higher education" in order to avoid federal enforcement proceedings (Texas Higher Education Coordinating Board, 1997). Since Governor William Clements issued the first plan, each subsequent governor has submitted a follow-up designed to continue increasing minority representation in Texas higher education. The most recent plan, Access and Equity 2000, sought increases reflecting the proportion of college-age minorities in the Texas population as well as increased minority representation on faculties and advisory boards. Similarly, until affirmative action was terminated in California, the state of California engaged in numerous activities to increase minority representation on campus, ranging from academic preparation programs at community colleges to actively recruiting minority applicants to the campuses of the University of the California (UC).

Nonetheless, many of the admissions policies implementing affirmative action, particularly at the institutional level, have not been made explicit, and their effect in practice has often been unclear. As Kane noted, "Nearly two decades after the U.S. Supreme Court's 1978 *Bakke* decision, we know little about the true extent of affirmative action admissions by race or ethnicity.... Hard evidence has been difficult to obtain, primarily because many colleges guard their admissions practices closely" (Kane, 1998a, p. 17).

11

One recent study used survey data to estimate the extent in practice of race-based preferences in higher education admissions. Kane (1998b) estimated how race/ethnicity, high school grade-point average (GPA), scores on the Scholastic Assessment Test (SAT), participation in student government and athletics, and college selectivity affected the probability of acceptance to college for a nationally representative sample of the high school class of 1982. Kane found that, holding constant the factors other than race, black and Hispanic applicants had an appreciable advantage over white applicants, but only in selective colleges. In the most selective colleges (those in the top quintile of selectivity), Kane estimated that the average advantage of black applicants was "equivalent to nearly a full point increase in high school grade-point average (on a four-point scale), or to several hundred points on the SAT" (Kane, 1998b, p. 438). The data upon which Kane's estimates were based, however, are now quite dated, and the sample did not allow estimates specific to states or individual postsecondary systems. Thus research leaves unclear how substantial preferences were in the states that have been at the center of the debate about the elimination of affirmative action, such as California and Texas.

Recent Policies Curtailing Group-Based Preferences

Although this report focuses on California, recent initiatives curtailing affirmative action have been proposed or enacted in several states.

California

The first recent major rollback of affirmative action in higher education was the enactment of SP-1 in 1995 by the University of California Board of Regents. SP-1 stated that "the University of California shall not use race, religion, sex, color, ethnicity, or national origin as criteria for admission to the University or to any program of study." This resolution was a response to executive orders issued by Governor Pete Wilson that severely curtailed affirmative action in a broad range of state procurement and administrative decisions.

In November 1996, California voters approved Proposition 209, which eliminated the consideration of race, ethnicity, and gender in public employment, public contracting, and education. In effect, Proposition 209 provided constitutional backing for SP-1. The US Supreme Court refused to hear a challenge to Proposition 209 in November 1997, thus allowing the measure to stand. Admission decisions based on these new policies went into effect with students seeking admission for the Spring quarter 1997-1998. These actions, taken together, represent a full repeal of affirmative action policies in California's state system of higher education.

Texas

Around the same time, the Fifth Circuit Court of Appeals ended the use of any race-based consideration in admission decisions in the area under its jurisdiction. In 1992, four white students who had been denied entrance to the University of Texas law school filed suit against the university, claiming that the partially race-based admission process violated their Fourteenth Amendment rights. Four years later, the Fifth Circuit Court of Appeals upheld their claims. "The case against race-based preferences does not rest on the sterile assumption that American society is untouched or unaffected by the tragic oppression of its past. Rather, it is the very enormity of that tragedy that lends resolve to the desire never to repeat it, and find a legal order in which distinctions based on race shall have no place" (*Hopwood v. Texas*, (1996), as quoted in Feinberg, 1998, p.12).

The principle enunciated in *Hopwood* appears to be inconsistent with the standards enforced by the OCR investigations of the state's efforts to remedy the remaining vestiges of *de jure* segregation in public higher education. First in 1980 and again in 1987, the OCR found that Texas had not made adequate progress in eliminating such problems and required that additional plans be adopted in order to avoid federal action. Texas was informed by the OCR in 1997 (after the *Hopwood* ruling) that its higher education system would once again be reviewed to ensure that an OCR-approved plan had been effectively implemented and that all traces of segregation had been eliminated in compliance with Supreme Court precedent (Siegel, 1998). The standard for the OCR review was set by *United States v. Fordice*, a 1992 case in which the U.S. Supreme Court held "that any state with a history of segregation in higher education must implement affirmative measures, including racial preferences to eliminate those vestiges." This standard differs from the one set out in *Hopwood*, which allows the use of racial preferences, but only when a state entity is acting to remedy present effects of past discrimination at a specific institution (THECB, 1997). The OCR review is still in progress and it is uncertain how OCR standard in *Fordice* may affect the interpretation or implementation of *Hopwood*.

The *Hopwood* case has been returned to the Court of Appeals three times, and the most recent ruling suggests that race may play some role in admissions. In its most recent ruling, the Court of Appeals did not overturn the original *Hopwood* ruling, but it did rule that the District Court injunction prohibiting the University of Texas Law School from any use of race in making admissions decisions was overly broad and excessive. The case has been remanded again to Circuit Court for additional action. The degree to which racial preferences will be allowable in Texas and in other states under the jurisdiction of the Fifth Circuit therefore remains uncertain.

Washington

Initiative 200 (I-200) was passed by the voters of Washington state in November 1998. Like Proposition 209 in California, it restricts the use of race in employment, education, and contracting. In response to the initiative, University of Washington (UW) President Richard L. McCormick announced that UW would suspend the use of race, ethnicity, and gender in admissions beginning in Spring 1999. It is important to note, however, that I-200 was passed as a law, not as an amendment to the state constitution as was Proposition 209 in California. It is therefore still uncertain whether I-200 will supersede existing laws that allow the use of race in employment and contracting decisions. It does not, for example, apply to federally funded state programs that must comply with federal nondiscrimination laws.

Florida

The Board of Regents of the State University System of Florida voted in favor of Governor Jeb Bush's "One Florida" Initiative in February 2000. This plan eliminates the use of race as a factor in admission decisions in the Florida University system and outlines an alternative, race-neutral admission process. Florida planned to have its new admission criteria in effect for students graduating from high school in 2000.

Georgia

In Georgia, a case was filed in federal court by three white women denied admission to the University of Georgia. The plaintiffs sued the University and State Board of Regents under Title VI of the Civil Rights Act of 1964, alleging that they were discriminated against because of their race. In a decision handed down in July of 2000, federal judge B. Avant Edenfield of Georgia ruled that the 1978 *Bakke* decision has been misinterpreted and that diversity is "an amorphous, unquantifiable goal" that cannot be constitutionally justified. He nullified the University of Georgia's now-discarded policy of maintaining lower admission standards for blacks. In a non-binding opinion, he further criticized the university's use of race/ethnicity as a "plus" factor in the selection of 10 to 15% of the students in each year's entering class. The case is likely to be appealed to the 11th U.S. Circuit Court of Appeals (Denniston, 2000).

Michigan

Michigan has recently seen two challenges to affirmative action in higher education in federal court. The first, *Gratz v. Bollinger et. Al* (2000), challenged the use of race in admissions at the undergraduate level. The plaintiffs were unsuccessful applicants to the College of Literature, Science, and the Arts in Fall 1995 and Fall 1997, respectively. This case was recently decided in favor of the defendant. *Grutter v. Bollinger* (2001) was filed in 1997 against the University of Michigan's law school, challenging the use of race in its admission policy. In March, 2001, District Court Judge Bernard Friedman ruled that the law school's admissions policies considering race were unconstitutional and a violation of Title VI of the 1964 Civil Rights Act. Both cases have been certified as class actions and are expected to be appealed.

Policy Responses to Challenges to Affirmative Action

Policymakers in the university systems of Texas, California, and Florida have tried in various ways to maintain diversity in the face of legal restrictions on affirmative action. In all three states, individual campuses have tried to recruit at high schools whose students are traditionally underrepresented in the college population. At the system level, outreach in California and Texas has focused on informing the public about the race-neutral admissions policies and on assuring minorities that the higher education system is still hospitable.

All three of these states have also instituted "X% plans"—that is, policies that admit a certain percentage of graduating public high school seniors automatically to their university systems, primarily on the basis of students' academic ranks within their high schools. The Texas legislature passed House Bill No. 588, also known as the 10% rule, in May of 1997. The measure mandates that public or private high school students whose GPA places them in the top 10% of their graduating class be admitted automatically to "each general academic teaching institution" if they graduated within the previous two years and filed the appropriate applications on time. The act also stipulates that the governing board of each such institution will decide on an institutional basis whether to automatically admit any student in the top 25% of his or her graduating class, but not in the top 10%. The legislature also outlined factors other than academic achievement that institutions were to take into consideration when admitting the rest of their freshman classes—factors related primarily to socioeconomic status, geographic region, and uncommon hardship. The admission criteria for students not in the top 10% or 25% of their class were to be published in the academic catalogs and made available to the public not later than one year before the date when they were to take effect. Similarly, the factors used in awarding competitive fellowships and scholarships were to be made public. The act has applied to all admissions and scholarship awards since the Fall semester of 1998.

In California, the top 4% of graduating seniors from each public high school are now eligible for admission to a school in the UC system, although not necessarily to the campus of their choice. Each school sets its own admission standards based on system policies, but a student deemed eligible is guaranteed admission to at least one of the UC campuses. The 4% plan in California has been termed ELC, eligibility in the local context. In addition to graduating in the top 4% of their class, students must fulfill a minimum course requirement that specifies the number and level of courses to be taken in high school subject areas, and they must submit ACT or SAT scores if the institution of their choice requires test scores. The ELC 4% plan is expected to be in effect for freshmen applicants in Fall 2001.

13

Unlike Texas and California, Florida established an alternative policy concurrently with terminating of race-based admissions. Beginning with the class of 2000, the top 20% of graduating seniors from each public high school will automatically be admitted to the state university system under the Talented 20 Program. Because of enrollment caps, however, students are guaranteed admission only to one of the 10 Florida universities, not necessarily to their top choice. The "One Florida" plan also calls for an additional $20 million in need-based financial aid. Under this initiative, universities are asked to address the financial-aid needs of students admitted under the Talented 20 Program before those of other students.

Washington state is in the process of reviewing its policies in response to Initiative 200. The UW Board of Regents is considering a proposal to allow the use of race and gender as factors in choosing the recipients of privately funded scholarships. Applicants would undergo screening based on neutral factors such as merit and need. From among those who pass the screen, students would be matched with scholarships. The aim is to attract minority students to the UW system.

Initial Effects of Policy Responses

Only Texas has fully implemented its "X% rule" admission policy. Holley and Spencer (1999) found that in its first year, the 10% rule had no significant impact on the number of minority students enrolled as first-time freshmen at the University of Texas at Austin and Texas A&M University, the state's two flagship schools. Only eight more black students and one fewer Hispanic student enrolled at UT-Austin in 1998 than in 1997. At A&M, 19 more black students and 62 more Hispanic students enrolled in 1998. However, the results of the 10% rule were available only for the first year (academic year 1998) and may not be indicative of the long-term effects of the new program.

Although no data on the effects of the 4% rule on freshman enrollment patterns at the University of California schools are available, a simulation study assessing the potential effects was conducted by Saul Geiser of the UC Office of the President.[1] For all California public high-school graduates for whom SAT scores were available, Geiser (1998) calculated an Academic Index score—an 8,000-point scale that gives approximately equal weight to a student's high school GPA and SAT scores. Students were then ranked by Academic Index score within each high school, and those in several top percentiles were combined into a simulated UC eligibility pool. Geiser found that limiting admissions to only the top 4% of students within each high school would have a modest impact on the racial/ethnic composition of the admitted population. Of the total eligible pool, 31% would be white, 47% Asian, 14% Latino, 3% black, and 5% "other." However, selecting only the top 4% within schools would dramatically reduce the number of students eligible for the UC system. When the vacant admissions slots were filled by selecting the top 8.5% of students remaining statewide, the representation of black and Hispanic students decreased and the representation of Asian students increased. The total eligible pool would comprise 30% white, 53% Asian, 10% Latino, 2% black and 5% "other."

The Present Study

The study presented here focuses on the University of California system to explore how eliminating affirmative action in college admissions affects the diversity of the admitted student bodies. It examines the effects of changing from the system in force before SP-1 to a race-neutral admissions policy based solely on SAT scores and high school GPA. It examines both the impact on racial/ethnic composition (that is, the proportion of admitted students who are white, black, Hispanic, and Asian) and the impact on other aspects of diversity, including parental education, first language, and location and type of high school attended.

The study also explores the effects of using criteria other than race and ethnicity to increase the diversity of the student population. The primary aim of these analyses is to examine whether other socioeconomic and educational background variables can be used as a proxy for race in college admission decisions.

Data and Methods

This study focuses on the University of California system and uses data from several sources. The College Entrance Examination Board provided us with complete data files for all California students who took the SAT in 1995 through 1998. These data included SAT scores, information identifying students' high schools and the colleges to which they had their scores sent, and background data from the Student Descriptive Questionnaire (SDQ) that students fill out when registering for the SAT. The SDQ included self-reported information about grade-point averages, racial and ethnic identity, courses taken, activities, parents' income, parents' education, and languages spoken at home. Aggregate data for the UC system on actual acceptances and enrollments by race/ethnicity, and on the probability of admission to specific campuses as a function of SAT scores and GPA (independent of race/ethnicity), were taken from tabulations published by the UC Office of the President. Information on the characteristics of high schools was taken from two sources: US Department of Education's Common Core of Data Public Schools files and the California Department of Education DataQuest records. These were merged with student-level records using a matching of high school identifiers provided to us by the College Entrance Examination Board.

The analyses reported here had several stages. The first stage entailed creation of race-neutral admission models for each UC campus. These models estimated the probability of admission as a function of SAT scores and GPA on the basis of aggregate data from the UC office of the President. The eight UC campuses were grouped into three

categories based on the apparent selectivity of their admissions; within each category, one model from one campus was chosen to represent all of the schools in that category. This stage of the analysis is described in detail in Appendix A.

In the second stage, the three models chosen in the first stage were applied to the College Board data to estimate the racial/ethnic composition of the admitted pool under race-neutral admissions rules. These analyses used 1998 and 1995 data and were limited to students who attended high school in California at the time they took the SAT. The models did not predict acceptance or rejection for individual students; rather, they predicted the probability of admission for students in a given range of SAT scores and GPA. These probabilities were applied to counts of tested students in each range to obtain estimated counts of admitted students. The resulting estimates of racial/ethnic composition were compared with actual admission data from the class admitted in 1999 (after SP-1 and Proposition 209 had been implemented) to confirm their reasonableness.

The application and acceptance process in the UC system can be seen as a sequence of filters, which are described in the following section: (1) taking the SAT; (2) meeting the UC system eligibility criteria, based on SAT scores and GPA; (3) applying to a campus at a given level of selectivity; and (4) being admitted to that campus. Our models represented a simplified, race-neutral version of the fourth of these filters. We examined the effects of the four filters, individually and in various combinations, on the diversity of the surviving pool of students. For example, by removing the application filter, we estimated the racial/ethnic composition that would result if all students were successfully encouraged to apply to campuses at all three levels of selectivity. We also examined the effects of these filters on other characteristics of the admitted groups: whether high school students attended an urban, suburban, or rural high school, the type of high school attended (e.g., public, private, religiously affiliated), parents' level of education, and first language spoken at home. The results of these analyses were compared with actual admission data from three years: 1995 (to represent policy before the enactment of SP-1 and Proposition 209), 1999 (to represent full implementation of these policies), and 1997 (to represent the transitional period).

In the third stage, a number of alternative admission models were applied to gauge their effects on the diversity of the admitted student population. These models used both individual variables (such as parents' education) and school characteristics (such as the percentage of students receiving free or reduced-price lunch). As part of this stage, we replicated the analyses performed by the UC office of the President to model the effects of 4%, 6% and 12.5% admission policies to ensure that our data and methods were consistent with those used in that study.

Steps in the Selection Process

The selection process entails a series of filters that progressively winnow the applicant pool. The filters in our models are the following.

The decision to take the SAT (or ACT). Because admission to the University of California system requires that students take the SAT or ACT, those who fail to do so remove themselves from the pool of potential students. This filter is nearly universal among selective colleges and universities nationwide, although a few institutions (e.g., Bates and Bowdoin) either do not use it or make the submission of scores optional.

In California, 98% of students who apply to the University of California take the SAT (Geiser, 1998), and our models accordingly simplified this filter slightly by considering only whether students take the SAT, not either the SAT or the ACT.

University system eligibility. The University of California screens students for eligibility to the entire UC system. Ineligible students are for the most part ineligible for admission to any of the eight campuses. However, at the outset of the study period, each campus was allowed to allocate up to 6% of its slots to UC-ineligible students, and up to two-thirds of these slots could be used for admitting disadvantaged students (*1996 Guidelines for Implementation of University Policy on Undergraduate Admissions,* http://www.ucop.edu/sas/exguides.html).

UC system eligibility was based on three criteria. First, GPA and SAT-I scores were combined on a sliding scale to set minimum requirements. (SAT-I refers to the basic verbal and mathematics tests, while SAT-II refers to a number of optional, subject-matter tests.) For example, students with GPAs of at least 3.29 were UC-eligible as long as their combined SAT-I scores were at least 570, while students with GPAs of 3.0 were required to have a combined SAT-I score of at least 1270. Second, students were required to take a set of required courses. Third, students had to take three SAT-II tests, "including writing, mathematics Level 1 or Level 2, and one test in one of the following areas: English literature, foreign language, science, or social studies," although they were not required to attain a specific score on these tests (*Admission as a Freshman,* http://www.ucop.edu/pathways/impinfo/freshx.html).

Our models simplified system eligibility by applying the UC GPA and SAT-I criteria but not the UC requirements for specific courses or for taking SAT-II tests. Because the system-eligibility filter is specific to the University of California system, we conducted parallel analyses that excluded it.

Application to a campus at a given level of selectivity. Students who elect not to apply to an institution remove themselves from the pool of potential students. We lacked data on actual applications, but we did have a record of all

15

institutions to which each student had his or her SAT-I scores sent. We treated sending a score as a proxy for application, thus overestimating by a presumably small amount the actual number of applications. We established a flag indicating whether a student had sent scores to *any* of the campuses within three levels of selectivity (see Appendix A):

- High selectivity: Berkeley and UCLA
- Moderate selectivity: Irvine, Davis, Santa Barbara, and San Diego
- Low selectivity: Santa Cruz and Riverside.

Predicted admission based on GPA and SAT. The probability that each student will be accepted to a campus at a given level of selectivity was predicted using logistic regression models derived from published campus-level admission statistics (see Appendix A).[2] We refer to this as a race-neutral admissions model because the probabilities assigned to students were unaffected by race or ethnicity (or any characteristics other than SAT scores and GPA). These models could not predict admission or rejection for individual students; rather, they predicted the probability of admission for students within a given range of GPA and SAT-I scores. These probabilities were multiplied by the number of students in each range to yield a count of "admitted" students.

Limitations

This study is limited to students who attended high school in California and who took the SAT, and it examines only the impact of race-neutral admission decisions to University of California campuses. Analyses in other states might yield substantially different results. Because this study sorts campuses into three categories and uses one model from one campus within each category to represent all campuses in that category, the findings do not necessarily apply to individual campuses. Moreover, this study is based almost entirely on data collected during the late 1990s, and patterns of application, test-taking, and acceptance may change with time. Nonetheless, we expect that the findings generalize in broad stroke to numerous other state university systems.

This study was also limited by the type of data to which we had access. For example, we had no access to individual-level data about acceptance or rejection, and the aggregate data on admissions probabilities were not available separately by race/ethnicity. That lack precluded more refined and powerful analysis.

The Effects of Current and Race-Neutral Selection on Racial/Ethnic Composition

Estimates of the effects of selection policies on diversity are presented by the selectivity of the institutions, starting with the most highly selective.

Effects in Highly Selective Institutions

Admissions to highly selective institutions were modeled loosely on UCLA and Berkeley. As noted, while the application filter showed whether students applied to either of these two campuses, the regression model used for both was derived from Berkeley data. Use of the UCLA data would not have greatly changed the results (see Appendix A).

The first screen applied, students' decision whether to take the SAT, substantially decreased the percentage of Hispanic students and increased the percentage of Asian students. In 1998, 31% of California high school graduates, but only 19% of those taking the SAT, were Hispanic (Table 1). Conversely, 15% of graduates but 23% of SAT-takers were Asian. This screen, however, only slightly reduced the representation of black students, who constituted roughly 7% of both graduates and SAT-takers. The decision to take the SAT also slightly reduced the representation of white students, who constituted 45% of graduates and 42% of SAT-takers.

It is important to note, however, that eliminating this screen—that is, having all students take the SAT—would not fully eliminate its effects. The students who decide against taking the SAT are presumably lower-achieving on average than those who do take it. Thus if all students took the SAT, many of those who currently do not take it would fail to gain admission because of low scores; and if the SAT were no longer used in admissions, some would fail to gain admission because of weaker academic records.

The UC system eligibility screen had a very different effect: it reduced the percentage of black students substantially and the percentage of Hispanic students more modestly. In 1998, 7% of California students taking the SAT were black, in contrast to 4% of those system-eligible in terms of SAT scores and GPA (see italicized panel of Table 1.) Hispanics constituted 19% of all SAT-takers but 15% of those eligible. Applying the eligibility screen slightly increased the representation of whites and Asians.

Table 1
Racial/Ethnic Composition, Highly Selective Campuses:
Actual, and Estimated Using All Screens and SAT+GPA Admissions Model

	Asian, Asian-American, Pacific Islander	Black or African-American	Hispanic	White	Other	Decline to State
Graduates, 1998	15	7	31	45	1	NA
SAT-takers, 1998	22	7	19	42	6	3
UC eligible, 1998	*25*	*4*	*15*	*46*	*6*	*3*
Eligible and applied to high-selectivity school, 1998	*36*	*4*	*15*	*35*	*7*	*2*
Admitted by neutral model, 1998	*38*	*2*	*9*	*42*	*7*	*3*
1995 Admitted class	36	7	19	31	2	5
1997 Admitted class	38	6	15	33	2	6
1999 Admitted class	41	3	10	35	2	9

Note: Race/ethnicity is based on student self-reports for all rows except the "Graduates" row, which is based on reports by school administrators. Estimates are italicized; other numbers are actual counts. Percentages may not sum to totals because of the exclusion of American Indian students and rounding.
Sources: Estimates reflect NBETPP analysis; admission figures are published figures from UC (http://www.ucop.edu/pathways/infoctr/introuc/prof_engin.html); counts of SAT-takers are based on NBETPP tabulations of data provided by the College Board; counts of graduates are from California Department of Education, Educational Demographics Unit (http://data1.cde.gov/dataquest).

Using the application screen—that is, dropping all students who did not apply to Berkeley or UCLA—did not affect the representation of Hispanics or blacks ("Eligible and applied to high selectivity school" row of Table 1). The *number* of minority students dropped by nearly half when this filter was applied, but that decrease was similar to the decrease in the total number of students in the pool. The application filter did, however, increase the percentage of Asian students and decrease the percentage of whites. Assuming that most of the students who requested that scores be sent to a particular campus actually applied for admission to that campus, it appears that Asians are particularly likely and whites less likely to apply to Berkeley and UCLA.

The final screen, the race-neutral admissions function based only on SAT and GPA, markedly reduced the representation of Hispanics and more still that of blacks. Black students dropped from 4% to 2% of the pool at this stage (see "Admitted by neutral model" row of Table 1), while Hispanics dropped from 15% to 9%. The offsetting increase was among white students, not Asians.

These screens have a cumulative effect, progressively reducing the representation of Hispanic and black students in the pool. That effect for Hispanics can be seen in Figure 1, which graphically represents the percentages in Table 1. The second bar shows a dramatic reduction from all graduates to those who took the SAT. The next screen, UC eligibility (simplified, as noted earlier, to reflect only SAT scores and GPA), produced a more modest but still appreciable drop. Application to high-selectivity schools had no effect on the representation of Hispanics, but the race-neutral admissions model reduced it substantially.
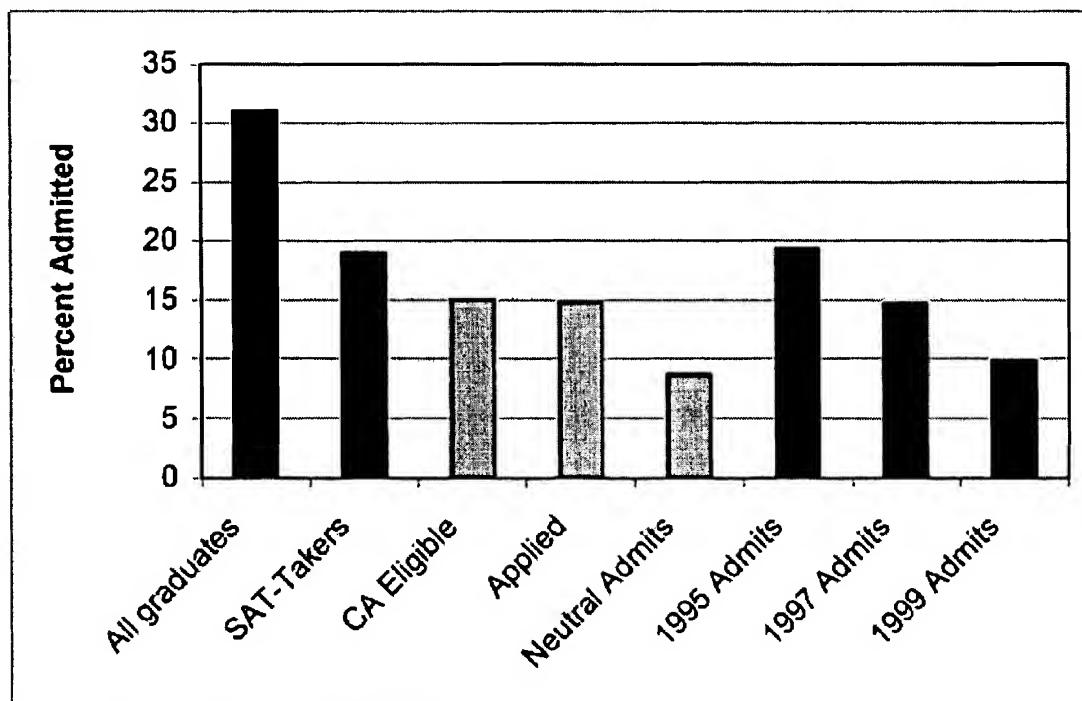
Figure 1. Hispanics as Percentage of Group Admitted, Highly Selective Campuses

In 1995, before the implementation of SP-1 or Proposition 209, Hispanics constituted 19% of students admitted to Berkeley and UCLA, which was almost exactly equal to their representation in the population of SAT-takers (Figure 1). Hispanics were thus overrepresented slightly relative to their numbers among UC-eligible students and substantially relative to a race-neutral policy. By 1999, after SP-1 and Proposition 209 were implemented, the representation of Hispanics admitted at these two campuses fell to roughly the percentage predicted by our race-neutral model. As Karabel (1998) noted, the admission of minorities fell after the enactment but before the implementation of SP-1 and Proposition 209 (note the drop between 1995 and 1997 in Figure 1).

The cumulative effects of these admissions screens on blacks present a somewhat different picture. While the sharpest drops in the representation of Hispanics arose from self-selection to take the SAT and the use of a race-neutral model, the declines for blacks arose primarily from the UC system eligibility screen and the race-neutral model (Figure 2). In 1995, blacks constituted 7% of students admitted to Berkeley or UCLA, almost exactly matching their representation among SAT-takers and high school graduates (Figure 2). Their representation among students actually admitted to Berkeley or UCLA, like that of Hispanics, dropped in both 1997 and 1999. Their representation among actual admissions in 1999, however, while very low, was substantially higher than was predicted by our simple GPA- and SAT-based race-neutral admissions model.
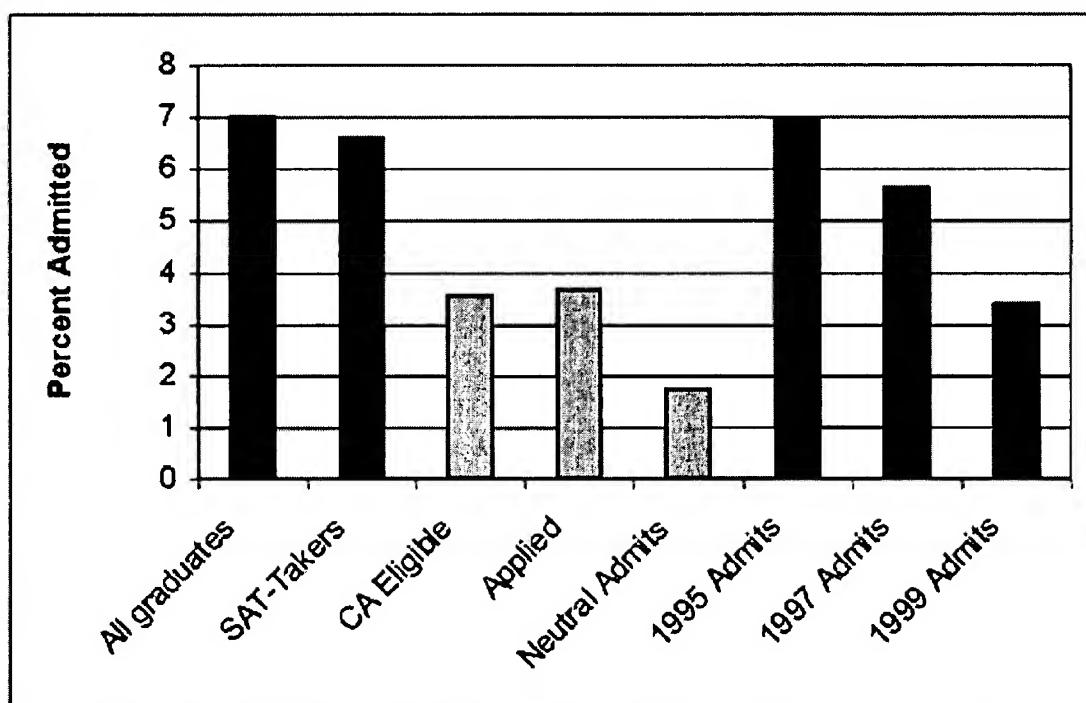
Figure 2. Blacks as Percentage of Group Admitted, Highly Selective Campuses

Our model did not match as well the representation of Asian and white students in the admitted pool. In 1999, 41% of the students admitted to Berkeley or UCLA were Asian, and 35% were white. Our race-neutral model predicted slightly fewer Asians (38%) and appreciably more whites (42%) than were actually admitted. We suspect but cannot verify that this is due to differences in the proportions of white and Asian students applying to selective private institutions in California and to colleges outside the state.

Because most states lack the system-eligibility screen used in California, we tested the generality of these findings. We applied the race-neutral admissions model based on SAT scores and GPA to all students who had sought admission to either UCLA or Berkeley and eliminated the UC system eligibility screen. Dropping that screen had almost no effect on the racial/ethnic composition of the group "admitted by neutral model" presented in Table 1. Recall that our simplified system eligibility rule is based solely on SAT scores and GPA, and the admissions model for the highly selective campuses applies such stringent requirements for those scores that the system eligibility screen is simply irrelevant.

In principle, one simple way to address the underrepresentation of minority students would be to encourage all students

18

to apply to the highly selective campuses. Therefore, in a second simplification, we applied the race-neutral admissions model to all students who took the SAT, regardless of system eligibility and of the schools to which they had their SAT scores sent. This too affected the racial/ethnic composition of the "admitted" pool only slightly. The total number of "admitted" students went up by more than half; the increases in the numbers of "admitted" blacks and Hispanics, however, were roughly proportional to that overall increase.

Some observers have argued that admissions tests such as the SAT should be abandoned in order to produce a student body more nearly representative of the racial/ethnic composition of the entire population. For example, in 1997, a university task force recommended that the University of California drop the SAT as an admission requirement to avoid a precipitous decline in the enrollment of minorities at the university's flagship campuses (Fletcher, 1997). Thus we estimated a second set of race-neutral models, based solely on GPA with no consideration of SAT, to explore how that criterion would affect the diversity of the accepted student population. (See Appendix A.)

A race-neutral model based solely on GPA creates a substantial underrepresentation of both black and Hispanic students, though a somewhat less severe one than that generated by the race-neutral model based on SAT and GPA together. This is shown in Table 2, which presents the racial/ethnic composition of the groups admitted by the SAT+GPA and GPA-only models. The second panel in Table 2, "SAT+GPA," presents estimates (discussed earlier) based on the SAT+GPA race-neutral model.[3] For example, using the SAT+GPA model with students who applied to either Berkeley or UCLA, 9% of those "admitted" would be Hispanic. In contrast, 13% of students selected by the GPA-only model would be Hispanic (third panel of Table 2). Thus the GPA-only model increased by roughly half the percentage of admitted students who are Hispanic; but even with this model, Hispanics were substantially underrepresented relative to the 31% of graduates and 19% of SAT-takers who were Hispanic. Similarly, using a GPA-only model increased the percentage of the admitted group who are black by about half, but would still admit less than half as many blacks as there were either among graduates or among SAT-takers.

Table 2
Racial/Ethnic Composition, Highly Selective Campuses, Estimated Using SAT+GPA and GPA-Only Admissions Models

	Asian, Asian-American, Pacific Islander	Black or African-American	Hispanic	White	Other	Decline to State
Graduates, 1998	15	7	31	45	1	NA
SAT-takers, 1998	22	7	19	42	6	3
SAT+GPA						
Applied and admitted	38	2	9	42	7	3
GPA only						
Applied and admitted	36	3	13	37	7	2

Note: Race/ethnicity is based on student self-reports for all rows except the "Graduates" row, which is based on reports by school administrators. Estimates are italicized; other numbers are actual counts. Percentages may not sum to totals because of the exclusion of American Indian students and rounding.
Sources: Estimates reflect NBETPP analysis; admission figures are published figures from UC
(http://www.ucop.edu/pathways/infoctr/introuc/prof_engin.html); counts of SAT-takers are based on NBETPP tabulations of data provided by the College Board; counts of graduates are from California Department of Education, Educational Demographics Unit (http://data1.cde.gov/dataquest).

The improved representation of minority students achieved by using a GPA-only model rather than a GPA+SAT model, however, would come at a price. Grading standards are inconsistent from high school to high school, and there is evidence that they vary across types of school. For example, grading tends to be more lenient in schools with high poverty rates (U.S. Department of Education, 1994). Absent a measure standardized across schools, these inconsistencies would introduce additional arbitrariness into the admission process and could lower the overall level of academic preparedness of the admitted group.

Effects in Moderately Selective Institutions

We classified as moderately selective the campuses at Irvine, Davis, Santa Barbara, and San Diego. Our race-neutral admissions model for these campuses was based on data from Irvine.

The first two screens applied in examining admissions to moderately selective institutions were the same as for highly selective schools—that is, deciding to take the SAT and meeting UC system eligibility requirements. Thus, the representation of non-Asian minority students fell substantially before use of the filter of application to a moderately selective campus: Hispanic representation was reduced by the SAT-taking screen, and black representation by the UC system eligibility screen.

19

Although the self-selection of eligible students to apply to moderately selective campuses reduced the pool by about 40%, it affected the racial/ethnic composition of the pool only modestly (Table 3). Hispanic students constituted 15% of the eligible pool but 13 % of the eligible students who applied to such an institution. Black students, who constituted a meager 4% of eligible students, made up only 3% of those who were eligible and applied. The representation of Asian-American students increased appreciably with this filter, and the representation of whites dropped.

Table 3
Racial/Ethnic Composition, Moderately Selective Campuses: Actual, and Estimated Using All Screens and SAT+GPA Admissions Model

	Asian, Asian-American, Pacific Islander	Black or African-American	Hispanic	White	Other	Decline to State
Graduates, 1998	15	7	31	45	1	NA
SAT-takers, 1998	22	7	19	42	6	3
UC eligible, 1998	*25*	*4*	*15*	*46*	*6*	*3*
Eligible and applied to moderately selective school, 1998	*32*	*3*	*13*	*42*	*6*	*2*
Admitted by neutral model, 1998	*33*	*2*	*12*	*43*	*7*	*2*
1995 Admitted class	35	3	14	42	4	2
1997 Admitted class	36	3	13	44	3	5
1999 Admitted class	36	2	11	41	2	8

Note 1: Race/ethnicity is based on student self-reports for all rows except the "Graduates" row, which is based on reports by school administrators. Estimates are italicized; other numbers are actual counts. Percentages may not sum to totals because of the exclusion of American Indian students and rounding.
Note 2: An earlier version of this table was corrected on January 17, 2002.
Sources: Estimates reflect NBETPP analysis; admission figures are published figures from UC (http://www.ucop.edu/pathways/infoctr/introuc/prof_engin.html); counts of SAT-takers are based on NBETPP tabulations of data provided by the College Board; counts of graduates are from California Department of Education, Educational Demographics Unit (http://data1.cde.gov/dataquest).

When used with UC-eligible students who applied to one of the four moderately selective institutions, the race-neutral admissions model had little effect on the racial/ethnic composition of the student pool. The number of students fell by roughly one-fourth with use of the admissions model, but the reduction was nearly proportional to the racial/ethnic groups. The percentage of Hispanics decreased only from 13% to 12%; that of blacks dropped from 2.6% to 2.2 %.

For moderately selective campuses as well, our models suggest that by 1999 admissions in all four campuses taken together were largely race-neutral. The composition of the group admitted in 1999 was very similar to that predicted by our race-neutral model (Table 3; compare the "Admitted by neutral model" row with the actual figures for 1999 admissions).

However, at the moderate-selectivity campuses taken together—in contrast to the highly selective campuses—the composition of the classes admitted changed only modestly from 1995 to 1999. Between 1997 and 1999, the percentage of admitted students who were black declined from 3% to 2%, and the percentage who were Hispanic from 14% to 11% (Table 3). These small changes after affirmative action was terminated suggest that racial/ethnic preferences had been much less substantial at the moderately selective campuses, taken together, than at the highly selective campuses.

We again examined the effects of removing the UC eligibility requirements and the application screen. Removing the former only slightly increased the size of the pool of students, and had only trivial effects on the ethnic composition of the accepted student group. In other words, for the most part students who were not UC system eligible either did not apply to any of these colleges or were predicted to be rejected by our admissions model. This was due mainly to students who took the SAT but did not apply to any of the four institutions. Removing the application screen—in effect, having all students who took the SAT apply—increased the number of students "accepted" by more than half. This increase, however, was roughly proportional to racial/ethnic groups, and so would raise the percentage of students who were black or Hispanic only slightly.

Effects in the Least Selective Institutions

In many respects, admission to the least and the moderately selective UC campuses was similar. In both cases, the main reduction in the representation of Hispanics occurred though the self-selection of students to take the SAT (Table 4). The UC system eligibility screen brought a modest further reduction, but the application screen and the race-neutral model had little effect. In contrast, blacks were proportionately represented among SAT-takers, and the primary reduction in the representation resulted from the application of the UC eligibility screen.

Table 4
Racial/Ethnic Composition, Least Selective Campuses: Actual, and Estimated Using all Screens and SAT+GPA Admissions Model

	Asian, Asian-American, Pacific Islander	Black or African-American	Hispanic	White	Other	Decline to State
Graduates, 1998	15	7	31	45	1	NA
SAT-takers, 1998	22	7	19	42	6	3
UC eligible, 1998	*25*	*4*	*15*	*46*	*6*	*3*
Eligible and applied to low-selectivity school, 1998	*33*	*3*	*16*	*38*	*7*	*2*
Admitted by neutral model, 1998	*33*	*3*	*15*	*39*	*7*	*3*
1995 Admitted class	32	4	19	38	4	2
1997 Admitted class	35	3	15	39	3	5
1999 Admitted class	33	3	15	39	2	7

Note: Race/ethnicity is based on student self-reports for all rows except the "Graduates" row, which is based on reports by school administrators. Estimates are italicized; other numbers are actual counts. Percentages may not sum to totals because of the exclusion of American Indian students and rounding.
Sources: Estimates reflect NBETPP analysis; admission figures are published figures from UC (http://www.ucop.edu/pathways/infoctr/introuc/prof_engin.html); counts of SAT-takers are based on NBETPP tabulations of data provided by the College Board; counts of graduates are from California Department of Education, Educational Demographics Unit (http://data1.cde.gov/dataquest).

Admission of black and Hispanic students to the least selective campuses changed little from 1995 to 1999 and matched our race-neutral model reasonably closely in all years. This suggests that students' preferences played little role in admission to these institutions.

The Effects of Admission Filters and Race-Neutral Selection on Other Aspects of Diversity

The diversity of the student body has numerous aspects in addition to race and ethnicity. In this section, we examine the effect of each filter in the admission process on other aspects of diversity: the geographic location and type of the secondary schools students attended, the education level of students' parents, and the languages students speak at home.

Information on these variables was obtained from the Student Descriptive Questionnaire (SDQ) that students complete when registering for the SAT and thus is subject to the errors common to survey data of this sort. For example, students may not consistently characterize the language used in their homes. Because the effects on racial/ethnic diversity are largest at the highly selective campuses, these analyses are limited to them.

Geographic Location

The SDQ offers six options for classifying the location of students' high schools: large city, medium city, small city, suburban, rural, and other. In 1998, 30% of SAT takers in California reported attending a secondary school located in a large city, and 60% in cities of all sizes (Table 5).[4] Thirty percent attended school in a suburban area, and only 5% in a rural area.

Table 5
Geographic Composition, Highly Selective Campuses, Using System Eligibility, Application, and SAT+GPA Admissions Model

	Large City	Medium City	Small City	Suburban	Rural	Other
SAT-takers, 1998	30	16	14	30	5	5
UC Eligible, 1998	*28*	*17*	*15*	*32*	*6*	*3*
Eligible and applied to high-selectivity schools, 1998	*32*	*16*	*12*	*35*	*3*	*2*
Admitted by neutral model, 1998	*28*	*16*	*12*	*40*	*3*	*1*

Although the effects of the application filters on geographic representation are modest compared with those on racial/ethnic composition, they do somewhat increase the percentage of students who are from suburban schools. All

21

three filters—UC system eligibility, application to a highly selective campus, and predicted admission—contribute to this effect; taken together, they increased the representation of suburban students from 30% of all SAT-takers to 40% of those admitted by a race-neutral model (Table 5). This increase was offset by smaller decreases in the percentages from schools in other locations. Surprisingly, the admission filter had only very small and inconsistent effects on the representation of students from large cities.

School Type

The SDQ allowed students to specify four types of high school: public school, religiously affiliated school, independent school without religious affiliation, and other. Of California students who took the SAT in 1998, 81% attended a public school, 13% attended religiously affiliated schools, 2% attended a non-religious independent school, and 4% attended alternative types of school (Table 6).

Table 6
School Type, Highly Selective Campuses, Using System Eligibility, Application, and SAT+GPA Admission Model

	Public	Independent	Religious	Other
SAT-takers, 1998	81	2	13	4
UC Eligible, 1998	82	3	13	3
Eligible and applied to high-selectivity schools, 1998	83	3	13	1
Admitted by neutral model, 1998	83	4	12	1

The effects of the admission filters on the mix of school types were minor. At all stages of selection, between 81% and 83% of students were from public schools, and 12% or 13% were from religious schools (Table 6). The filters reduced the representation of students from "other" schools and increased that of students from non-religious independent schools, but both of these groups constituted only a small percentage of the total group at each stage.

Parents' Education

Students were asked to report the highest education level attained by their fathers and mothers. The SDQ offered the following response options: grade school, some high school, high school diploma, business school, some college, associate's degree, BA degree, some graduate school, and graduate degree. We collapsed these nine categories into five:

- No high school diploma;
- High school diploma;
- Some higher education;
- College degree (BA degree);
- Beyond BA.

Each of the filters increased the representation of students whose parents had at least a bachelor's degree. The three filters taken together increased the representation of children of college-educated mothers by 50% or more (Table 7). For example, 18% of SAT-takers but 30% of "admitted" students had mothers with more than a BA degree. The two categories of mothers with a BA or beyond were roughly of equal size and showed approximately the same effects. While all three screens contributed to this pattern, use of the race-neutral admissions model had the largest effect. These increases were offset by decreases in the representation of children of the three categories of less-educated mothers, with proportionately the greatest reduction occurring for children of mothers with no high school diploma.

Table 7
Mother's Education, Highly Selective Campuses, Using System Eligibility, Application, and SAT+GPA Admissions Model

	No HS Diploma	HS Diploma	Some Higher Ed	BA Degree	Beyond BA
SAT-takers, 1998	14	16	34	19	18
UC Eligible, 1998	11	14	32	21	21
Eligible and applied to high-selectivity schools, 1998	12	13	28	23	23
Admitted by neutral model, 1998	6	11	25	28	30

The same general pattern appeared with fathers' education, for students who reported that as well, although some specific effects were different. The admission filters—particularly the race-neutral admission model—had more impact on the representation of children whose fathers had post-graduate education. Students who reported that their fathers had more than a BA constituted 24% of all SAT-takers but 43% of "admitted" students (Table 8).

Table 8
Father's Education, Highly Selective Campuses, Using System Eligibility, Application, and SAT+GPA Admissions Model

	No HS Diploma	HS Diploma	Some Higher Ed	BA Degree	Beyond BA
SAT-takers, 1998	13	14	28	20	24
UC Eligible, 1998	11	12	26	22	29
Eligible and applied to high, 1998	11	10	22	23	33
Admitted by neutral model, 1998	6	7	19	25	43

Home Language

Students responding to the SDQ were given three options for describing the first language they speak at home: only English, English and another language, and another language. As Table 9 indicates, most students who took the SAT in 1998 spoke only English at home; 21% of the SAT takers primarily spoke a language other than English; and 16% spoke a combination of English and another language.

The effects of the admission filters on the representation of these three groups were inconsistent. For example, use of the UC system eligibility filter slightly decreased the representation of students who speak other languages at home, from 21% to 19%; use of the application filter increased their representation to 25%; and use of the race-neutral admission model reduced it again to precisely the level—21%—shown among all SAT-takers (Table 9). We suspect that these effects stem from the fact that the categories "English and other language" and "another language" include some of the Asian students who are overrepresented at the most selective UC campuses.

Table 9
Representation of Home Languages, Highly Selective Campuses, Using System Eligibility, Application, and SAT+GPA Admission Model

	English Only	English & Other Language	Another Language
SAT-takers, 1998	63	16	21
UC Eligible, 1998	65	16	19
Eligible and applied to high, 1998	54	21	25
Admitted by neutral model, 1998	59	20	21

The Effects of Alternative Selection Models on Diversity

The "X%" policies adopted by California, Texas, and Florida are intended to capitalize on the unequal distribution of low-scoring minority students among high schools in order to increase the representation of racial/ethnic minorities in the pool of admitted students. An alternative approach to that end is to give weight to other demographic variables that may act as a proxy for race/ethnicity. In this section, the effects of both approaches are examined.

Top X % Policies and Other Aspects of Diversity

As summarized above, Geiser (1998) simulated the impact four different X% policies would have on both the racial/ethnic composition and on the average academic preparedness of groups admitted to the University of California. Here, we examine the effect of these policies on other measures of ethnic diversity as well. Unlike Geiser (1998), however, we ranked students solely on their GPAs, an approach that is more consistent with the X% policies actually implemented to date.[5]

We modeled four X% rules. The first ranked all students in public high schools statewide by their GPAs and admitted the top 12.5%. The second ranked all students within each school and admitted the top 12.5% from each school.[6] The third rule admitted the top 6% within each school, and the fourth admitted the top 4% within each school. To yield an admitted group of students that represents 12.5% of graduating public school students, the third and fourth models also accept the top 6.5% and 8.5% of students statewide after removing the top 6% and 4% from within each school,

respectively.

The baseline rule is to attract the top 12.5% across the state. Automatically accepting the top 4% from each school before accepting the remaining top 8.5% would not appreciably affect the academic qualifications of admitted students overall or the proportion of black or Hispanic students. Accepting the top 6% within each school would likewise have little effect on diversity and academic qualifications, but it would reduce the mean SAT of accepted black and Hispanic students appreciably relative to the baseline, by 45 and 33 points respectively.

Table 10
Modeled Results of Top 4%, 6% and 12.5% Admission Policies

	White	Asian	Black	Hispanic	Other	Total
Top 12.5% Across State (Baseline)						
% of Admitted Group	49	29	2	10	10	
Mean SAT	1211	1222	1136	1126	1210	1204
Mean GPA	3.87	3.89	3.90	3.93	3.88	3.89
Top 4% Within HS						
% of Admitted Group	49	29	2	10	10	
Mean SAT	1214	1223	1117	1115	1210	1204
Mean GPA	3.88	3.90	3.87	3.93	3.88	3.89
Top 6% Within HS						
% of Admitted Group	48	29	2	11	10	
Mean SAT	1216	1221	1091	1093	1208	1200
Mean GPA	3.89	3.90	3.84	3.89	3.88	3.89
Top 12.5% Within HS						
% of Admitted Group	42	27	4	18	9	
Mean SAT	1198	1173	1001	999	1168	1145
Mean GPA	3.87	3.88	3.60	3.68	3.85	3.83

In contrast, accepting the top 12.5% within each school would have dramatic effects compared with the baseline condition of accepting the top 12.5% statewide. The percentage of admitted students who are black would double, from 2% to 4%, and the percentage of Hispanic students would increase from 10% to 18%. This increase in diversity, however, would occur at the cost of a large drop in the academic qualifications of admitted minority students. The mean SAT scores of black and Hispanic students would drop 135 and 127 points, respectively, and their mean GPAs would drop by .3 and .25, respectively. The academic qualifications of the total admitted pool would drop as well, although less markedly. The mean SAT would drop 59 points relative to the baseline rule.

Accepting the top 12.5% within each school rather than the top 12.5% across the state to would have similar effects on other aspects of diversity. Table 11 shows that moving from 12.5% statewide to accepting the top 4% or 6% within schools would have little impact on the distribution of geographic location, first language, or education level of a student's mother. However, accepting the top 12.5% within schools increases the representation of urban students, students who speak a language other than English at home, and students whose mothers have limited education.

Table 11
Impact of Top X% Policies on Other Aspects of Diversity

	12.5% Across State	4% Within School	6% Within School	12.5% Within School
Location				
Urban	.55	.56	.56	.63
Suburban	.39	.38	.38	.30
Rural	.06	.06	.06	.07
Home Language				
English	.65	.65	.64	.60
Eng. and Other Language	.15	.15	.15	.16
Other than English	.19	.19	.20	.22
Mother's Education				

Some Higher Ed.	.30	.30	.30	.30
College Degree	.23	.23	.23	.19
Beyond BA	.23	.23	.23	.18
Missing	.03	.03	.03	.03

Giving Preference to Other Aspects of Diversity

The race-neutral admission models presented above result in an over-representation of white and Asian students. The model for highly selective schools also results in an overrepresentation of suburban students, students who speak only English at home, and students whose parents are highly educated. In this section, we explore the impact on racial/ethnic composition of giving preference to students who are from low-income families, whose mothers have little education, who attended high school in urban or rural areas, who attended high schools with low graduation rates, or who attended high schools in which a high percentage of students received free or reduced-price lunch.[7]

In each of these analyses, we awarded the equivalent of a 200-point SAT bonus to students who came from the most disadvantaged background in terms of one of these variables. The preference awarded for each step on a variable depended on the number of categories the variable had. For example, in this analysis, mother's education had only two categories (BA or beyond versus no BA), while income was broken into 14 categories (Table 12). Accordingly, while all students whose mothers lacked a BA were given the full 200-point preference, each decrease of one step on the income variable warranted an additional 1/14 of the total 200 points, or roughly 17 points per step.

Table 12
Variables, Number of Levels, and Preference per Step Applied in Alternative Models

Variable	Levels	Effective SAT Point Boost Per Level
Income	14	16.67 per step
Mother's Education	BA or Beyond No BA	0 200
HS Graduation Rate	>75% 50 to 75% <50%	0 100 200
Free/Reduced Lunch	Continuous, from 0% to 95%	2.1 for each 1 percent increase in percent free lunch
School Location	Suburban/Other Urban/Rural	0 200

Table 13 displays the effect of giving preference to each variable, first individually and then in combination. When combinations of variables were used, the maximum impact of the combination was set to 200 points. Since school-level data were available only for public schools, these analyses were run on a reduced data set. The first two rows of Table 13 compare the results of the SAT and GPA-only models and show that the reduced and full data set yielded about the same ethnic mix of admitted students.

Giving preference to students based on any of the variables decreases the representation of Asian students and increases that of white, black, and Hispanic students. Giving preference to students from schools that have low graduation rates and that are located in either urban or rural settings has the largest impact on the representation of white students and the smallest on that of Hispanic and black students. However, giving preference to students whose mothers are less well-educated or whose families are poor has the largest impact on the representation of Hispanic and black students.

Perhaps most important, even the largest effects of giving preference based on demographic variables do not come close to making the representation of black and Hispanic students in the admitted groups proportionate to their numbers in the pool of potential students. As Table 13 shows, even giving preference to urban and rural students with low family incomes, whose mothers have not completed college, and who attend high schools with low graduation rates and high free and reduced lunch rates still results in a dramatic underrepresentation of black and Hispanic students and an overrepresentation of white students.

Table 13
Students Admitted (%)

Modeled Variable	Asian	Black	Hispanic	White	Other	Decline to state

SAT + GPA Full Sample	37.5	1.7	8.8	41.7	7.5	2.8
SAT + GPA Alt. Sample	39.9	1.8	8.3	40.0	7.5	2.4
SAT + GPA + Income	32.6	2.3	11.2	46.0	7.1	0.9
SAT + GPA + Mother's Education	30.0	2.2	11.7	47.9	7.0	1.3
SAT + GPA + Location	29.2	2.2	10.3	48.6	7.1	2.5
SAT + GPA + Graduation Rate	30.5	1.9	9.1	48.8	7.1	2.6
SAT + GPA + Free Lunch	30.6	2.1	10.2	47.4	7.0	2.5
SAT + GPA + Income + Free Lunch + Mother's Ed + Location	30.2	2.1	10.4	48.8	7.1	1.4

Discussion

Our analyses addressed two broad questions: what stages of the admissions process produce the under representation of minorities, and what effects might different admissions processes—including both a strict race-neutral policy and alternative preferences based on variables other than race and ethnicity—have on the diversity of the student population? Investigating these general questions shed light on several others as well: the extent of racial/ethnic preferences in place before the end of affirmative action, the relationships between preferences and the selectivity of campuses, and the effects of alternative admissions policies on the characteristics of admitted students, both minority and non-minority.

Replacing the former admissions process that included preferences with a race-neutral model based solely on GPA and SAT scores had major effects at the two most selective campuses in the UC system but much smaller effects at both moderate- and low-selectivity campuses. Kane (1998b) found a similar pattern in national data, but the present analyses show that this pattern is maintained even within a single university system in which some admissions criteria are common across campuses. Both black and Hispanic populations were also noticeably underrepresented in our moderately and least selective environments, but this under representation stemmed primarily from factors other than the actual admissions process—in particular, whether a student decided to take the SAT and whether the student met the minimum eligibility criteria for the UC system. Once students passed these two hurdles, the actual admission decision had a substantial impact on the representation of black and Hispanic students only for highly selective campuses.

The adverse effects of a race-neutral admissions process were complex. An under-representation of Hispanics (but not of blacks) arose because of the large percentage of the former group who decided not to take the SAT. (Because we lacked scores for those students, we could not estimate how many would have been admitted had they taken the test.) Scores had an adverse impact at two stages, the UC eligibility stage and the campus-level race-neutral admissions process. However, these effects were in some ways duplicative, and eliminating the UC eligibility screen had little impact on the composition of the groups admitted to highly selective campuses if the campus-level admissions process remained unaltered. The decision to apply to selective campuses had little impact on diversity; the race-neutral admissions model would have produced a similar mix of students even if all students who had taken the SAT had applied.

The adverse impact of a race-neutral admissions policy was not solely the result of group differences in scores on admissions tests. A race-neutral model based solely on GPA also produced an under-representation of minorities, albeit a less severe one. The effects of using GPA alone are smaller because the gap between groups in grades is smaller than the gap in average scores. The reasons for this difference and the potential effects of relying more on GPA, however, remain uncertain.

None of the alternative admissions models we analyzed could replicate the composition of the student population that was in place before the termination of affirmative action in California. Giving preference to students on the basis of other socioeconomic or demographic variables had only modest effects on the representation of black and Hispanic students; none that we examined brought minority students to proportional representation. Some of these preferences, however, increased the representation of whites at the cost of Asians. Guaranteeing admission to top students within each school—the "X-percent rules"—would substantially increase the representation of minority students only if the percentage within each high school guaranteed admission is large. Of the models we examined, only admitting the top 12.5 percent of students from each high school—in effect, basing admission to the UC system solely on rank within high schools—led to a large increase in the representation of black and Hispanic students. Applying the 12.5% rule, however, had a large cost: it caused a sizable drop in both average SAT scores and average GPA, and that decline was particularly large for black and Hispanic students. As Geiser (1998, p. 4) noted, "Redefining the UC eligibility pool to include the top 12.5% of each school would, in short, produce a bifurcated eligibility pool with severe academic disparities along racial/ethnic lines."

Admissions systems differ greatly, and the UC system studied has elements not shared by many others—in particular, the dual screening, first for UC system eligibility and subsequently for admission to a specific campus. Moreover, the effects of preferences and other admissions policies depend on the characteristics of the student populations from

which universities draw. For example, the Hispanic population in Florida is unlike that in California in several important respects, and the effects of admissions policies on access for Hispanic students therefore could be substantially different in Florida.

Nonetheless, we expect that many of the basic conclusions we reached in examining the California system will apply at a general level to many university systems nationwide because group differences in prior academic performance and test scores are typically large. The task of providing access to postsecondary education for underrepresented minorities without frank preferences is likely to be difficult and complex throughout the nation. Many of the alternatives will have unintended effects, such as lowering the average level of qualification among admitted minority students. However, some important details of the impact of alternative admissions policies will vary from one system and population to another. Therefore, it would be prudent to examine proposed alternatives carefully before implementation and to monitor their effects once implemented in order to maximize their positive effects and minimize unintended outcomes.

Notes

This research was conducted under the auspices of the National Board on Educational Testing and Public Policy (NBETPP), and is published here with the Board's permission. The NBETPP is located in the Lynch School of Education at Boston College and is an independent body created to monitor assessment in American education. The NBETPP provides research-based information for policy decision making, with special attention to groups historically underserved by the educational system. In particular, the Board a) monitors testing programs, policies, and products; b) evaluates the benefits and costs of specific testing policies; and c) evaluates to what extent professional standards for test development and use are met in specific contexts.

[1] Note that Geiser's simulation modeled a policy that differs from the actual 4% policy implemented in California. Rather than selecting the top 4% of students from within each high school based on their GPAs, Geiser based selection on students' combined high school GPA and SAT scores.

[2] More precisely, the models were weighted least squares regressions of logits of admissions probabilities on GPA and SAT. This is equivalent except in estimation method to a logistic regression of the probability of admission on GPA and SAT. See Appendix A.

[3] The estimates in Table 2, unlike those in Table 1, do not use the UC system-eligibility screen. That screen is based in part on SAT scores; applying it here would therefore not provide a clear contrast between admissions models that do and do not use SAT scores. The estimates in Table 2 reflect only students who have taken the SAT, however, as we have data for only those students.

[4] Because could not locate data that describe the distribution of all high school graduates in terms of school location, school type, or household language use, we were compelled to use student self-reports from the SDQ for this information. Therefore, the tables in this section consider no groups larger than the pool of SAT-takers and lack the "Graduates" rows that appear in the tables in the previous section.

[5] Geiser calculated an "Academic Index score" using the following formula: $AI = 1,000GPA + 2.5SAT$. Geiser then ranked students within schools based on AI.

[6] To determine the top 12.5 % within a graduating class, we used data from the CDE to establish the number of students who graduated and multiplied this by .125. Students were then ranked by their GPA within schools, and the number of students representing the top 12.5% within the graduating class were admitted. The same procedure was repeated for the 6% and 4% models.

[7] These analyses roughly follow the approach Wightman (1997) took in examining the result of giving preference to students with disadvantaged backgrounds for law school decisions.

[8] Because they are simpler to interpret, we also estimated linear probability models. As expected, however, they were problematic. In some cases, they yielded considerably weaker fits, gave impossible estimates for some cells, and showed inappropriate residuals.

[9] Values of 0 and 1 were set to .001 and .999, respectively, to calculate logits.

[10] In the case of Berkeley, the interactive model predicted somewhat better than the non-interactive model, but nonetheless yielded unreasonable estimates for some cells. In the case of Irvine and Santa Cruz, the interaction term added little to prediction.

References

27

Bakke v. Regents of the University of California, 553 P. 2d 1152 (Cal., 1976).

Denniston, L. (2000, July 28). Judge dismisses interpretation of high court's Bakke decision Anti- affirmative action ruling could be next test for justices' current view. *The Baltimore Sun*, p. 3A.

Feinberg, W. (1998). *On higher ground: Education and the case for affirmative action.* Columbia University, New York: Teachers College Press.

Fletcher, M. A. (1997). UC may drop SAT entry requirement; California system fears sharp decline in Black, Latino enrollment. *The Washington Post*, September 20, p. A3.

Holley, D. and Spencer, D. (Winter, 1999). The Texas 10 percent plan. *Harvard Civil Rights-Civil Liberties Law Review.*

Geiser, S. (1998). *Redefining UC's Eligibility Pool to Include a Percentage of Students from Each High School: Summary of Simulation Results.* Oakland, CA: University of California Office of the President.

Gratz and Hamacher v. Bollinger and the University of Michigan, 97-CV-75231-DT (2000).

Grutter v. Bollinger and the University of Michigan, Civil Action # 97-75928.

Hopwood v. Texas, 78 F. 3d 932, 946 (5th Cir. 1996).

Kane, T. J. (1998a). Misconceptions in the debate over affirmative action in college admissions. In G. Orfield and E. Miller (Eds.). *Chilling Admissions: The Affirmative Action Crisis and the Search for Alternatives.* Cambridge, MA: Harvard Education Publishing Group, pp. 17-32.

Kane, T. J. (1998b). Racial and ethnic preferences in college admissions. In C. Jencks and M Phillips, *The Black-White Test Score Gap*, Brookings, pp. 431-456.

Karabel, J. (1998). No alternative: The effects of color-blind admissions in California. In G. Orfield and E. Miller (Eds.). *Chilling Admissions: The Affirmative Action Crisis and the Search for Alternatives.* Cambridge, MA: Harvard Education Publishing Group, pp. 33-50.

Siegel, R. (1998). The racial rhetorics of colorblind constitutionalism: The case of *Hopwood v. Texas.* In Post, R. and Rogin, M. (Eds.), *Race and representation: Affirmative action.* New York: Zone Books.

Texas Higher Education Coordinating Board. (1997). *Alternative diversity criteria: Analyses and recommendations.* Austin, Texas: Texas Higher Education Coordinating Board.

Tracy v. Board of Regents. CV 497-45 (2000).

United States v. Fordice, No. 90-1205 (1992).

U.S. Department of Education (1994). *What do student grades mean? Differences across schools.* Research Report (January).

University of California Board of Regents v. Bakke, 438 U.S. 265 (1978).

Wightman, L. F. (1997). The threat to diversity in legal education: An empirical analysis of he consequences of abandoning race as a factor in law school admission decisions. *New York University Law Review*, 72(1), 1-53.

About the Authors

Daniel Koretz
Harvard Graduate School of Education
415 Gutman Library
6 Appian Way
Cambridge, MA 02138

Phone: (617) 384-8090
Fax: (617) 496-3095

E-mail: daniel_koretz@harvard.edu

Daniel Koretz is a professor at the Harvard Graduate School of Education and Associate Director of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). His work focuses primarily on educational assessment and recently has included studies of the validity of gains in high-stakes testing programs, the effects of testing programs on schooling, the assessment of students with disabilities, and international differences in the influences on mathematics achievement.

Michael Russell is a senior research associate with the National Board on Educational Testing and Public Policy and a Research Professor at the Center for the Study of Testing, Evaluation and Educational Policy at Boston College. His research interests include educational testing and technology.
E-mail: Mike Russell russelmh@bc.edu

Chingwei David Shin is currently a doctoral student enrolled in the Educational Statistics and Measurement program at University of Iowa. His research interests include validity issues in alternative assessments, item response theory, equating, scaling, and statistical methodology.
E-mail: ching-wei-shin@uiowa.edu

Cathy Horn is a research associate with the Civil Rights Project at Harvard University. She is currently working on a project assessing the effects of race-neutral admissions policies on diversity at public universities.
E-mail: chorn@law.harvard.edu

Kelly Shasby Burling is working on her PhD in Educational Research, Measurement, and Evaluation at Boston College. She is currently working at Harvard University's Center for International Development within the John F. Kennedy School of Government on projects using technology in educational settings to promote sustainable development in developing countries.

Appendix A
Constructing Race-Neutral Admissions Models

The baseline admission models presented in this report are intended to reflect some of the most important characteristics of the University of California system, but it was not possible to match that system precisely. Unlike Bowen and Bok (1998), we lacked data from campuses on the characteristics of individual students, and we lacked important aggregate variables separately by race, such as the probability of acceptance given SAT scores and HSGPA. We lacked campus-level data on the characteristics of the relatively few students admitted despite failing to meet the UC eligibility requirements ("admitted by exception"). More important, we lacked campus-level information on the more numerous UC-eligible students who were admitted for reasons other than only their academic performance, variously measured. Diverse factors, including personal disadvantage and school characteristics, can be used in deciding whether to admit these students, who can constitute 25% to 50% of an admitted class (http://www.ucop.edu/pathways/infoctr/introuc/select.html).

Even without this information, however, it was possible to create an approximation to admission in the UC system as it would operate without racial preferences. The steps we followed are presented here.

The starting point for our baseline models was data showing the numbers of total and accepted applicants by SAT score and HSGPA, separately by campus, for all programs except Engineering. In the data we obtained, SAT scores were broken into five ranges, and HSGPA was broken into six. Table A.1 shows the data we obtained for one of the campuses; these were not further broken into racial/ethnic categories. We analyzed the probability of admission in each cell of this matrix—that is, the ratio of the number admitted to the total number of applicants—separately for each of the eight UC campuses.

Table A.1
Admissions Probabilities, Berkeley
All Programs Except Engineering, 1999

GPA	SAT Composite Score					
	490 - 790	800 - 990	1000 - 1190	1200 - 1390	1400 - 1600	Overall
2.82 - 2.99			149 / 10	115 / 4	16 / 0	280 / 14
			6.70%	3.50%	6.70%	5.00%
3.00 - 3.29	61 / 6	288 / 23	831 / 65	730 / 48	138 / 9	2048 / 151
	9.80%	8.00%	7.80%	6.60%	6.50%	7.40%
3.30 - 3.59	65 / 11	408 / 33	1336 / 94	1620 / 116	423 / 45	3852 / 299

	16.90%	8.10%	7.00%	7.20%	10.60%	7.80%
3.60 - 3.89	52 / 4	421 / 57	1726 / 175	3025 / 414	830 / 259	6054 / 909
	7.70%	13.50%	10.10%	13.70%	31.20%	15.00%
3.90 - 3.99	01-May	59 / 14	353 / 48	798 / 181	198 / 107	1413 / 351
	20.00%	23.70%	13.60%	22.70%	54.00%	24.80%
4	21 / 5	210 / 59	1673 / 506	4775 / 2100	3179 / 2405	9858 / 5075
	23.80%	28.10%	30.20%	44.00%	75.70%	51.50%
Overall	204 / 27	1386 / 186	6068 / 898	11063 / 2863	4784 / 2825	25796 / 7072
	13.20%	13.40%	14.80%	25.90%	59.10%	27.40%

SOURCE: University of California, Office of the President
(www.ucop.edu/pathways/infoctr/introuc/prof_except.html).

As expected, these data show tremendous differences in the selectivity of the UC campuses. The probability of admission to Berkeley along with UCLA, the most selective, is low for all applicants other than those with both high SAT scores and high HSGPA (Figure A.1). For those with low SAT scores, only very high grades increase the probability of admission at all. Similarly, only very high SAT scores help students with grades as low as B (3.0), and even high scores do not increase the probability of admission greatly.
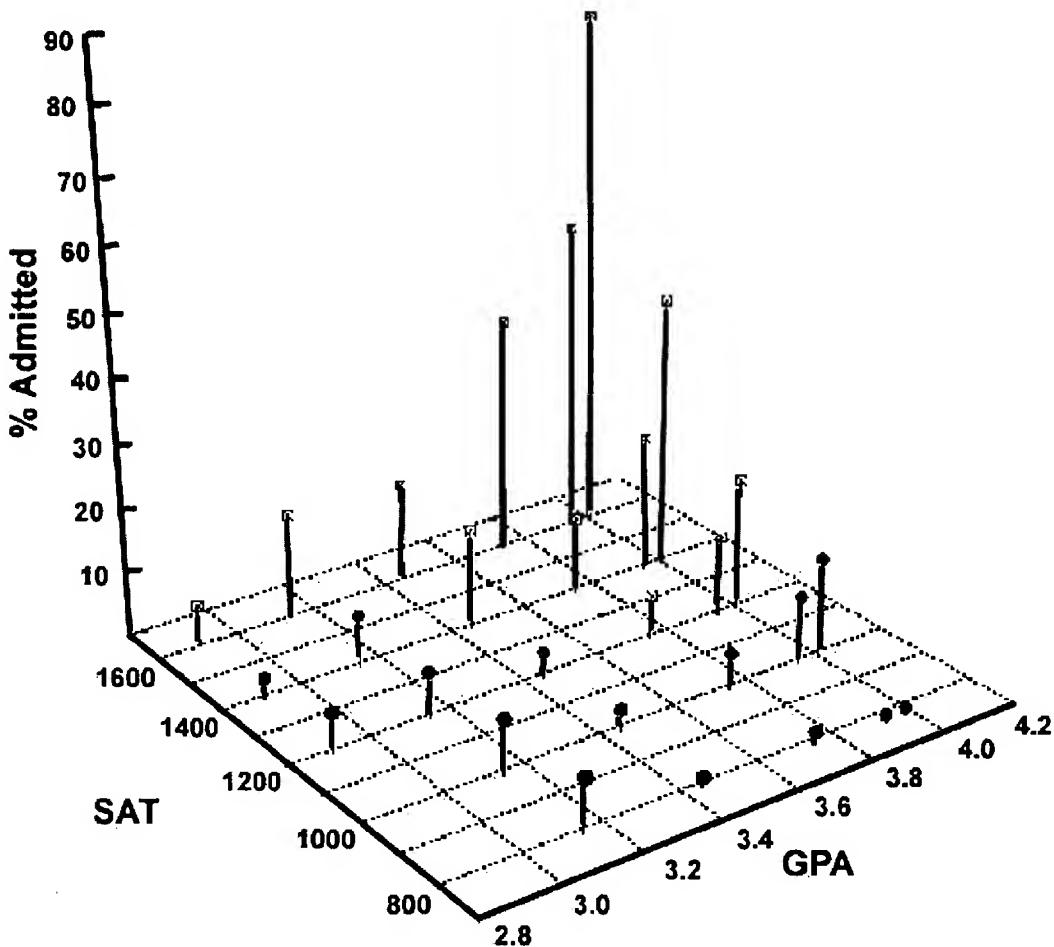

Figure A.1. Probability of Admission to Berkeley, by SAT and GPA

Riverside presents a dramatically different picture(Figure A.2). Most students with either high GPA or high SAT scores are admitted, and acceptance rates are above 50% for most groups in the graph. Note that as a result, the relationship between admissions probabilities and both SAT scores and grades is relatively weak; lines drawn between most points in planes parallel to either the GPA or the SAT axis in Figure A.2 have shallow slopes.
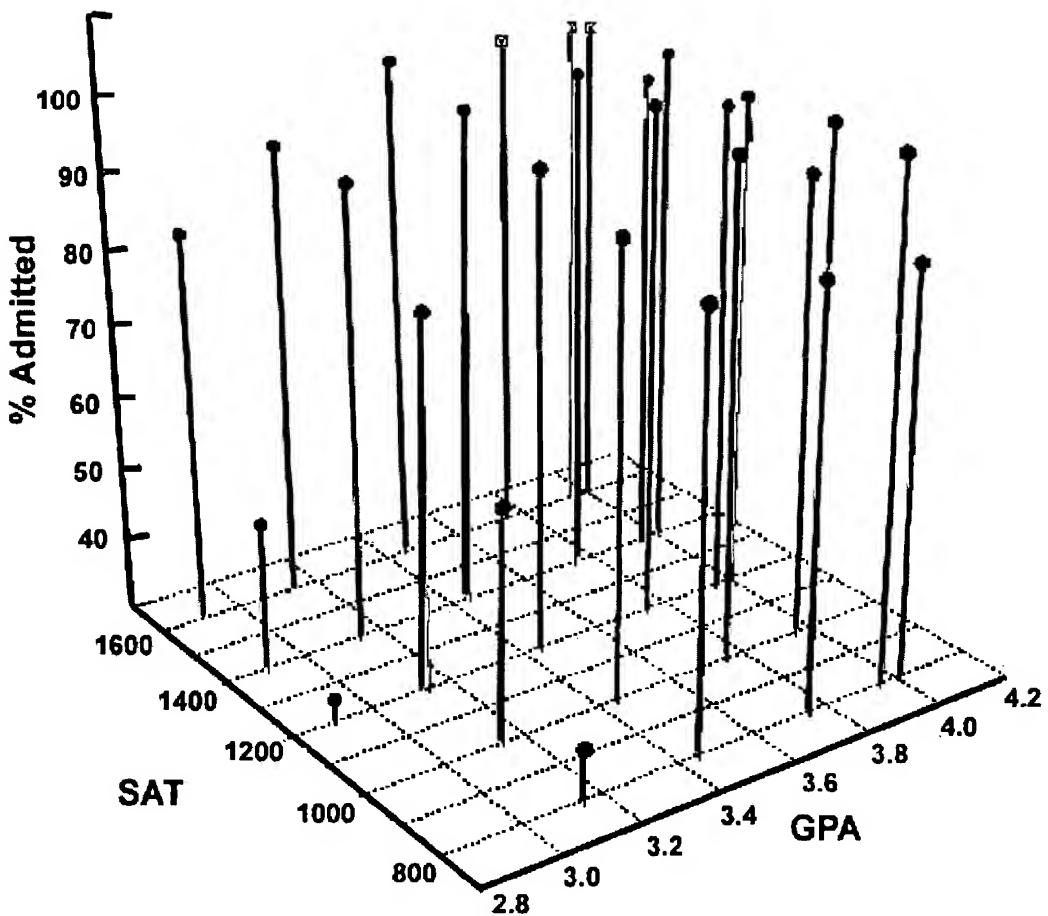
Figure A.2. Probability of Admission to Riverside, by SAT and GPA

To estimate the probability of admission for individual students, we fit models estimating admission probability as a function of SAT scores and GPA, separately for each campus. Given the dichotomous outcome and the distribution of probabilities (the number of cells with either very high or very low probability), we used logistic models to estimate both non-interactive and interactive models. All models were weighted by the number of applicants in each cell.[8] Because all of the variables were categorical, the logistic models could be estimated as ordinary least squares models by taking the logits of the probabilities for each cell[9]:

$$y = \ln\left[\frac{p}{(1-p)}\right] = \alpha + \beta_1 SAT + \beta_2 GPA + \varepsilon \qquad (1)$$

$$y = \ln\left[\frac{p}{(1-p)}\right] = \alpha + \beta_1 SAT + \beta_2 GPA + \beta_3 (SAT \cdot GPA) + \varepsilon \qquad (2)$$

These are equivalent to logistic probability models but are simpler to estimate. For example, model 1 is equivalent to:

$$p = \frac{1}{1 + \exp(-\alpha - \beta_1 SAT - \beta_2 GPA - \varepsilon)}$$

These simple logistic models fit the data closely. The R^2 values for the non-interactive models, adjusted for

shrinkage, were all greater than or equal to .79, and six of the eight were greater than or equal to .90. The interaction added appreciably to the fit in the ease of Berkeley and UCLA but had little impact elsewhere.

Examination of the data and the models suggested that the UC campuses fell into the following three levels of selectivity.

- High selectivity. Berkeley and UCLA were clearly more selective than any of the other campuses. Although the models for these two schools had substantially different parameter estimates, the probabilities they predicted were very similar.
- Moderate selectivity. This group includes four schools: Irvine, Davis, Santa Barbara, and San Diego. They appeared to place somewhat different weights on GPA and SAT scores. Irvine and San Diego showed greater effects of GPA than did Davis and Santa Barbara. Santa Barbara and San Diego showed stronger effects of SAT scores than did Irvine or, especially, Davis. As a group, however, they were distinct from the high- and low-selectivity schools.
- Low selectivity. Santa Cruz and Riverside appeared to be the least selective of the eight campuses. They gave similarly little weight to low SATs; Riverside gave less weight to low GPAs.

These three groups are the basis for our high-, moderate-, and low-selectivity scenarios. The high-selectivity scenario was based on the Berkeley campus. The mid-selectivity scenario was based on the Irvine campus, and the low-selectivity scenario was based on the Santa Cruz campus. The non-interactive logistic model was used in all cases.[10]

Because SAT scores and GPA are correlated in the UC data, we used separate regression models to estimate the effects of selection models based on solely GPA or SAT rather than both. These models were simply:

$$y = \ln\left[\frac{p}{(1-p)}\right] = \alpha + \beta \cdot SAT + \varepsilon$$

$$y = \ln\left[\frac{p}{(1-p)}\right] = \alpha + \beta \cdot GPA + \varepsilon$$

These were estimated using marginal percentages in the data tables such as Table 1.

Using data from the College Board, these models were applied to records of all California high school seniors who took the SAT in 1995 and 1998. Students' SAT-Total and HSGPA were used to place them in cells corresponding to the UC admissions probability matrix, and on that basis each student was assigned a probability of admission to a campus at each of the three levels. The models estimated logits, so the estimated probabilities were simply the anti-logits of the model estimates, that is:

$$\hat{p} = \frac{\exp(\hat{y})}{1 + \exp(\hat{y})}$$

These probabilities were multiplied by the counts in each cell and rounded to get counts of "admitted" students. For certain purposes, the counts provided by each model were adjusted to approximate total admissions for all of the campuses (either two or four) at that level of selectivity, but in most cases, only the characteristics of the "admitted" group (i.e., the percentage of admitted students who were black) were used.

A series of additional flags was created for each student in the College Board database. The data contain no information about actual applications but do include the identities of all schools to which students had their SAT scores sent. Students who sent scores to any of the UC campuses were assumed to have applied to that campus. Four flags were created in this way: sent scores to any campus; sent scores to one of the two high-selectivity campuses; sent scores to any moderate-selectivity campus; and sent scores to any of the two low-selectivity schools. These were treated as application flags but may overestimate applications, presumably by a modest amount. An additional eligibility flag was created using the UC systemwide eligibility criteria for SAT scores and GPA. The UC requirement that SAT-II scores be submitted was not used in creating this flag. All of these flags were set to 0 when the condition was not met and to 1 when the condition was met.

Applying these screens and our baseline admission models in various combinations allowed us to examine the effects of various stages of the admission process and to simulate the effects of alternatives on the composition of the accepted group. For example, removing the application flag provides an upper-bound estimate of the effects of efforts to encourage all UC-eligible students to apply to all campuses; removing the eligibility screen estimates the impact of moving to a system in which students apply directly to UC campuses without first being filtered by a systemwide eligibility screen.

32 BEST COPY AVAILABLE

Appendix B
Comparison of Merged and Full Databases

As noted in the body of this report, data on high school characteristics were unavailable for many of the California students for whom we had data from the College Entrance Examination Board. When school characteristics were not needed, we used the full database, but analyses involving any school characteristics were necessarily conducted with a reduced database. This Appendix briefly describes the two databases.

The full database was defined as all students in the College Board database who had valid data on GPA (variable = RECUMGPA) and SAT scores (variable = SATTOTAL). In 1998, that selection criterion left a total of 131,406 out of 152,680 students in the College Board California data. A key school variable obtained from the California Department of Education (Educational Demographics Unit, http://data1.cde.gov/dataquest) was counts of students in grade 12. We were able to merge this variable into the records of 93,027 students who also had valid data on SAT scores and GPA. These 93,027 students are represented in the merged database. Thus, merging school data caused a loss of 29% of the full database.

Although this sample loss was large, tabulations suggest that it did not materially affect our analyses. Table B1 provides a comparison of the racial/ethnic composition of the full and merged databases at four stages of the process of admission to the highly selective campuses. In all cases, the percentages are similar. More important for our purposes, the change in percentages caused by each of the filters is similar in the merged and full databases. The conclusions presented in the paper would not differ greatly if one of these databases were substituted for the other.

Table B1
Racial/Ethnic Composition, Merged and Full Databases
Model Based on SAT+GPA (Row Percents)

	Asian, Pacific Islander	Black or African-American	Hispanic	White	Other	Decline to State
Merged Data						
SAT-takers, 1998	24.2	6.6	19.5	40.2	5.6	2.8
UC eligible, 1998	27.4	3.5	15.3	44.6	5.9	2.4
Eligible and applied to high selectivity schools, 1998	38.4	3.5	14.7	33.9	6.7	2.1
Admitted by neutral model, 1998	39.9	1.8	8.3	40.0	6.9	2.4
Full Database Without Merge						
SAT-takers, 1998	22.4	6.6	19.0	41.9	5.8	3.2
UC eligible, 1998	25.3	3.6	15.1	46.1	6.2	2.9
Eligible and applied to high selectivity schools, 1998	36.0	3.7	14.8	35.4	6.9	2.4
Admitted by neutral model, 1998	37.5	1.7	8.8	41.7	7.0	2.8

BEST COPY AVAILABLE   **34**

OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gscisucla.edu

35

Education Policy Analysis Archives

State-Mandated Testing and Teachers' Beliefs and Practice

Sandra Cimbricz
Canisius College

Abstract
In this article, I examine the relationship between state-mandated testing and teachers' beliefs and practice. The studies reviewed suggest that while state testing does matter by influencing what teachers say and do, so, too, do other things, such as teachers' knowledge of subject matter, their approaches to teaching, their views of learning, and the amalgam of experience and status they possess in the school organization. As a result, the influence state-mandated testing has (or not) on teachers and teaching would seem to depend on how teachers interpret state testing and use it to guide their actions. Moreover, the influence state testing may or may not have on teachers and teaching expands beyond individual perceptions and actions to include the network of constructed meanings and significance extant within particular educational contexts. Consequently, although a relationship between the state-mandated testing and teachers' beliefs and practice does exist, testing does not appear to be an exclusive or a primary lever of change. There is further suggestion that it might not be a substantive one either. Studies that provide a richer, more in-depth understanding of the relationship between state-mandated testing and teaching in actual school settings, therefore, not only point toward important directions for future research in this area, but are greatly needed.

Introduction

reported by teachers in Grant's (2000) study were negative, however. Several teachers, especially elementary and high school math and English teachers, for example, cited greater collaboration with their peers. The development of informal networks and relationships, therefore, was reported as one of the key benefits stemming from the changes made in the state-mandated testing program.

Zancanella (1992) investigated the influence of the language arts segment of the low stakes, state-mandated Missouri Mastery and Achievement Tests given in grades 6, 7, 8, and 9 on the thought and action of three, tenured, middle school/junior high school English-language arts teachers. These case studies revealed that the import that the new state tests held for these teacher's thinking and practice in the teaching of literature was related to two factors: 1) the fit between the teacher's preferred approach to teaching literature and the conceptions of literature embodied in the state tests and 2) the amount of "curricular power" the teacher held—that is, the teacher's place in the curricular decision-making structure of the school. He concludes that teachers' responses to state policy were cast in terms of their prior learning, beliefs, and attitudes. In other words, the biography of the teacher's own past experiences was an important force in the ways in which these three teachers responded, especially as they related to:

> (a) the degree to which teachers' conceptions of the subject matches the conception of the subject the tests represent, a version of what is often called "curricular alignment"; and (b) the amount of what might be called "curricular power" the teacher possesses, the amalgam of experience, status, and position in the school organization that determines how much say the teachers has in both formal and informal decisions about which ways of teaching a subject are viewed as legitimate. (p. 292)

Though tenured, Ms. Kelly, for example, reported feeling pressure from the school administration to have her students perform well on the tests and the need to change her teaching style to do so. Another teacher (who also taught in same school as Ms. Kelly) saw her inductive approach to literature and the nurturing of lifelong readers as being at odds with the state tests, but did not feel threatened that low test scores would damage her reputation or position. As department chair and a veteran teacher with thirteen years experience, Ms. Martin reported that she felt free to "go beyond quiet resistance to 'hammer away' at her principal" about issues related to the test (p. 228). The third teacher, Mr. Davidson, saw the tests as compatible with his idea of teaching literature, although there were times when the tests were seen as intrusive and counterproductive. Despite his misgivings, his way of teaching literature was seen as compatible, if not "aligned" with the new test. Mr. Davidson, therefore, did not see a need to adjust his teaching and believed issues related to the new test had little direct consequence on his teaching.

The studies reviewed in this section indicate that teachers' interpretations of state testing are influenced not only by state testing itself, but the particular beliefs, knowledge, and experience individual teachers possess. The view that teachers' practices are subject to multiple influences, therefore, emerges as key. According to Grant (1999), such a view fosters a richer understanding of teach decision-making, in that "[i]t obviates the notion that any single factor ... or set of factors ... substantially influence[s] teachers' practices. Instead, what teachers do in their classrooms is likely to be influenced by a range of factors reflecting a variety of sources" (p. 238).

Discussion

All of the studies reviewed consistently confirm that state-mandated testing does matter and does influence what teachers say and do. But while these studies suggest that the instructional methods teachers employ, the materials they use, and the activities they plan are, to some degree, shaped by the form and content of state-mandated tests and the state objectives that accompany them, there appears to be no clear or consistent pattern of influence. As such, the research that is currently available presents a picture more complicated than clear and begs further elucidation.

Some of the studies indicate that the effects of statewide testing vary according to the "stakes" involved. Because high-stakes tests and/or testing programs are used for important decisions, these tests are assumed to have more power than low stakes tests and/or testing programs to modify local behavior (Heubert & Hauser, 1998; Madaus, 1988). Following this line of argument, high-stakes test are more likely to impact, if not constrain, teachers' beliefs and practice. Brown (1992, 1993), Smith et al. (1989),

and Smith (1991), for example, argue that teachers from states with high-stakes state-mandated testing (e.g., Arizona, Illinois, New York, and Tennessee) reported and were observed tailoring their instructional methods, materials, and activities to the type of performance elicited by the state tests. Under these conditions, Brown, Smith et al, and Smith assert that the state tests became more the goal of instruction, rather than the means to assess it. In addition, these researchers contend that the attention the media, the state education department, and various people at the local level (e.g., administrators, principals, school board members, parents, and community members) pay to test scores may catapult state tests into even higher stakes status. In short, this camp of researchers argue that high-stakes state-level testing serves to constrain, if not homogenize instruction.

This, however, does not appear to be the case in Grant's (2001) study of two high school teachers in New York State—a state boastful of its high-stakes state tests known as the Regents Examinations. Grant suggests that while the influence of the state Regents exam was apparent, that influence held no privileged position and interacted with a range of other factors, particularly the teachers' views of subject matter and learners. Both teachers sought to prepare students for the same high-stakes test and yet, their instructional practices were found to be radically different. Muddying the waters further, Firestone, Mayrowetz, and Fairman's (1998) observations led them to conclude that the teaching of math was much the same in both Maine and Maryland despite the difference in stakes and state tests. What Firestone, Mayrowetz, and Fairman's work, in particular, reveals is that under some circumstances, state-mandated testing policies and the stakes attached to them can promote specific behavior and procedures in the classroom more easily than deeper understandings of subject matter and how to teach it. Taken together, Grant's (2001) and Firestone, Mayrowetz, and Fariman's (1998) studies importantly show that while the state test and the stakes involved may have influenced these teachers' practices in some way, these things did not necessarily construct or determine the instruction that was ultimately provided. Without denying that the effects of statewide testing could vary according to the "stakes" involved, these researchers suspect that the argument that high-stakes testing encourages teachers to teach to the test may be overrated. Instead, their work prompts important consideration of whether all high-stakes state-mandated testing is necessarily high-stakes for all—especially for those who teach in districts with large numbers of students failing state tests—and how teachers who are significantly influenced by state-mandated testing differ from those who are minimally influenced.

Other factors may also be interacting with the state test to influence teachers' beliefs and practice in addition to the stakes involved, teachers' subject matter knowledge, and teachers' views of learners and teaching. The grade level taught (e.g., elementary vs. secondary; grade levels tested vs. grade levels not tested) emerged as a potential factor (Glasnapp, Poggio, & Miller, 1991; Grant, 2000; Smith 1991) as did the amalgam of knowledge, beliefs, experience, status, and position a teacher possesses (Zancanella, 1992). District and building-level expectations, local context and conditions, and state and district policy climate also emerged as possibilities (Brown, 1992, 1993; Firestone, Mayrowetz, & Fairman, 1998; Grant, 2001; Smith, 1991; Smith et al., 1989). So while state-mandated testing may be an influence, it would appear to be one of many. Consequently, although a relationship between the state-mandated testing and teachers' beliefs and practice does exist, state-mandated testing does not appear to be an exclusive or primary lever of change.

Given what is currently known about the general relationship of state-level testing and teaching, one caveat must be offered. Research in this area overall tends to rely heavily on teachers' perceptions gained through surveys and interviews (e.g., Brown, 1992, 1993; Glasnapp, Poggio, & Miller, 1991; Grant, 2000). While it would be foolish to discount self-reported and interview data, the lack of observational data begs the question of how tests influence teaching in actual practice. Even when teachers report change, for instance, little is known about what that change looks like and whether change has occurred at all. Only a few studies that couple surveys and interviews with classroom observations currently exist (Firestone, Mayrowetz, & Fairman, 1998; Grant, 2001; Smith et al., 1989; Smith, 1991; Zancanella, 1992). While in-depth interviews may provide access to teacher perception, coupling interviews with classroom observation allows both thought and action to be put in context. Coupling interviews with observations appears to provide data not only on teachers' understandings of what and how to teach, but also on how those understandings are operationalized and carried out. That said, the question of "why" teachers changed or did not change in lieu of state-mandated testing begs further exploration. For these reasons, I believe that studies that allow for the contextualization of teachers' beliefs and practice hold considerable promise for future research. For if school reform via state-level testing is to prove

constructive for education, research on how teachers understand and interpret new policy in the context of their knowledge, beliefs, experience, and teaching circumstances is vital.

Conclusion

The studies reviewed suggest that while state testing does matter and influence what teachers say and do, so, too, do other things, such as teachers' knowledge of subject matter, their approaches to teaching, their views of learning, and the amalgam of experience and status they possess in the school organization. As a result, the influence state-mandated testing has (or not) on teachers and teaching would seem to depend on how teachers interpret state testing and use it to guide their action. Moreover, the influence state testing may or may not have on teachers and teaching expands beyond individual perceptions and actions to include the network of constructed meanings and significance extant within particular educational contexts. How tests matter then is not always clear and simple.

Given the limited number of studies that are currently available and the limited nature of the data on which many of these findings are based, studies that provide a richer, more in-depth understanding of the relationship between state-mandated testing and teaching in actual school settings not only point toward important directions for continued research in this area, but are greatly needed. For if state-mandated testing continues to be viewed as a viable mechanism of educational reform, it is necessary to understand the ways in which this mechanism is mediated through the local contexts and the minds, motives, and actions of teachers. Subsequently, studies that provide a richer, more in-depth understanding of the relationship between state-mandated testing and teaching in actual school settings not only point toward important directions for future research in this area, but are greatly needed.

Notes

1. Howe and Eisenhart (1990) provide five general standards for qualitative and quantitative educational research, specifically, that there be: 1) a fit between research questions and data collection and analysis techniques; 2) the effective application of specific data collection and analysis techniques; 3) an alertness to and coherence of background assumptions; 4) an overall warrant; and 5) an awareness of both external and internal value constraints.

2. Schwille et al., (1983) describes *teacher policy* "as the definitive allocation of public resources by working-level personnel in education" and *external policy* "as policy in the usual sense—the laws, regulations, and other directives of boards, legislatures, and executive departments" (p. 376). In this paper, external policy refers to state tests and/or a state-mandated testing policy.

3. I am referring to Smith et al.'s (1989) three key findings stated in the previous paragraph. Smith posits that 1) teachers were encouraged to use instructional methods and materials that resembled state-testing; 2) content areas not included on the state tests were neglected; and 3) time devoted to test preparation and testing reduced time available for other instruction. In her follow-up study, she expands this list from three to six.

4. Haney (2000) surveyed secondary math and English/Language Arts teachers, Hoffman et al. (1999) surveyed reading specialists statewide, and Gordon and Reese (1997) surveyed Texas teachers who were "graduate students in educational administration" (p. 349).

5. Exploration of instructional patterns in this study focused on the characteristics of mathematic lessons in terms of the "problem size" (i.e., large vs. small), "student activity" (i.e., practice vs. nonpractice), and "teacher activity" (i.e., tell procedure vs. develop concept) (Firestone, Mayrowetz, & Fairman, 1998, p. 104).

References

Barton, P. E., & Coley, R. J. (1994). *Testing in America's schools: Policy information report.* Princeton,

NJ: Educational Testing Service, Policy Information Center. (ERIC Document Reproduction Service No. Ed 366 616)

Bond, L.A., Baskamp, D. & Roeber, E. (1996) *The status report of the assessment programs in the United States*. Washington, DC: The Council of Chief State School Officers and Oak Brook, IL: North Central Regional Educational Laboratory.

Bond, L. A. & Cohen, D. A. (1991). The early impact of Indiana statewide testing for educational progress on local education agencies: Administrators' perceptions. In R. E. Stake (Ed.), *Advances in program evaluation: Vol. 1. Effects of mandated assessment on teaching* (pp. 75-99). Greenwich, CT: JAI Press Ltd.

Brown, D. F. (1992, April). *Altering curricula through state-mandated testing: Perceptions of teachers and principals*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Brown, D. F. (1993). *The political influence of state-mandated testing reform through the eyes of principals and teachers*. (ERIC Document Reproduction Service No. Ed 360 737)

Cohen, D., McLaughlin, M. W. & Talbert, J. (Eds.). (1993). *Teaching for understanding: Challenges for policy and practice*. San Francisco: Jossey-Bass.

Corbett, H. D. & Wilson, B. L. (1991). Two state minimum competency testing programs and their effects on curriculum and instruction. In R. E. Stake (Ed.), *Advances in program evaluation: Vol. 1. Effects of mandated assessment on teaching* (pp. 7-40). Greenwich, CT: JAI Press Ltd.

Ellwein, M.C., Glass, G. V., & Smith, M. L. (1988). Standards of competence: Propositions on the nature of testing reforms. *Educational Researcher, 17*, 4-9.

Feuer, M. J., & Fulton, K. (1993). The many faces of performance assessment. *Phi Delta Kappan, 74*, 478.

Firestone, W., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20* (2), 95-113.

Glasnapp, D. R., Poggio, J. P., & Miller, D. M. (1991). Impact of a "low stakes" state minimum competency testing program on policy, attitudes, and achievement. In R. E. Stake (Ed.), *Advances in program evaluation: Vol. 1. Effects of mandated assessment on teaching* (pp. 101-140). Greenwich, CT: JAI Press Ltd.

Gordon, S. P. & Reese, M. (1997). High stakes testing: Worth the price? *Journal of School Leadership. 7*, 345-368.

Grant, S. G. (1996). Locating authority over content and pedagogy: Cross-current influences on teachers' thinking and practice. *Theory and Research in Social Education, 24* (3), 237-272.

Grant, S. G. (2000, February). Teachers and tests: Exploring teachers' perceptions of changes in the New York State-mandated testing Program. *Education Policy Analysis Archives* [On-line serial], *8*(14). Available: http://epaa.asu.edu/epaa/v8n14.html

Grant, S. G. (2001). An uncertain lever: Exploring the influence of state-level testing on teaching social studies. *Teachers College Record, 103*(3), 398-426.

Haas, N. S., Haladyna, T. M. & Nolen, S. B. (1989). *Standardized testing in Arizona: Interview and*

written comments from teachers and administrators (Tech. Rep. No. 89-3). Phoenix, AZ: Arizona State University West Campus.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archieves* [On-line serial], *8*(41). Available: http://epaa.asu.edu/epaa/v8n41/

Haney, W., & Madaus, G. (1989) Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683-687.

Heubert, J. P. & Hauser, R. M. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Hoffman, J., Pennington, J., Assaf, L. & Paris, S. (1999). *High stakes testing in reading and its effects on teachers, teaching, and students: Today in Texas, tomorrow?* (Unpublished manuscript). University of Texas at Austin.

Howe, K. & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher, 19*(4), 2-9.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the NSSE, Part 1* (pp. 83-121). Chicago: University of Chicago Press.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1989). *A survey of Arizona teachers and school administrators on the uses and effects of standardized achievement testing* (Tech Rep. No. 89-2). Phoenix, AZ: Arizona State University West Campus.

Quality Counts 2000: Who Should Teach? (2000). *Education Week, XIX* (18*)*. [On-line]. Available: http://www.edweek.org/sreports/qc00.htm

Popham,W. (1987). Can high-stakes be developed at the local level? *NASSP Bulletin, 71*(496), 77-84.

Resnick, D. P. & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher, 14*(4), 5-20.

Romberg, T. A., Zarinnia, E. A., Williams, S. R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions.* Madison, WI: University of Wisconsin, Center for Educational Research, School of Education, and Office of Educational Research and Improvement of the United States Department of Education.

Rottenberg, C., & Smith, M. L. (1990, April). *Unintended effects of external testing in elementary schools.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Schwille, J., Porter, A., Belli, G., Floden, R. Freeman, D. Knappen, L., Kuhs, T., & Schmidt, W. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy* (pp. 370-391). New York: Longman.

Smith, M. L, Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1989). *The role of testing in elementary schools.* Los Angeles, CA: Center for Research on Educational Standards and Student Tests, Graduate School of Education, UCLA.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Stake, R. E. & Rugg, D. (1991). Impact on the classroom. In R. E. Stake (Ed.), *Advances in program evaluation: Vol. 1. Effects of mandated assessment on teaching* (pp. xix-xxii). Greenwich, CT: JAI Press Ltd.

Teacher Education Agency. (1997). *Texas student asssessment program technical digest for the academic year 1996-1997.* Austin, TX: TEA.

Wiggins, G. & McTighe, J. (1998). *Understanding by design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wiske, M. S. (Ed.). (1998). *Teaching for understanding: Linking research with practice.* San Francisco: Jossey-Bass.

Zancanella, D. (1992). The influence of state-mandated testing on teachers of literature. *Educational Evaluation and Policy Analysis, 14*(3), 283-295.

Acknowledgements

About the Author

Sandra Cimbricz
Teacher Education Department
Canisius College
2001 Main Street
Buffalo, NY 14208

Phone: 716-888-3731

Sandra Cimbricz recently earned her Ph.D. from the University at Buffalo and is currently an Assistant Professor of Teacher Education at Canisius College in Buffalo, New York. To contact her, please email:cimbrics@canisius.edu.

43

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e
Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

44

Socratic Pedagogy, Race, and Power:
From People to Propositions

Peter Boghossian
Portland State University

Abstract
Rud (1997) wrote in this journal: " Leaving aside the blatant (to my eyes at least) problems of
power and dominance of an elderly Greek citizen teaching a slave boy, this example [the Meno]
of teaching has always left me cold." Garlikov (1998) addressed Rud's criticism of the Socratic
dialogue. The present article addresses and extends Garlikov's response to cover general notions
of power, and shows how these may affect Socratic discourse. Socratic pedagogy is not merely
an illusory exercise where participants acquiesce to notions of truth because of power
differentials. But power relations play a role in all communicative contexts. However, in
Socractic pedagogy the adverse effects of power are greatly reduced and the focus is shifted from
people to propositions.

Introduction

The Meno has long been considered the paradigmatic example of the Socratic method. Here, solely by asking
questions, Socrates teaches a young slave boy that the area of a large square is twice the area of a smaller one. Some
scholars, however, find both the Socratic method generally, and this example specifically, to be problematic because
of notions of power and the influence this may have on the participants' responses. Garlikov engaged part of the
criticism that relates to the idea of respondents being logically led to given conclusions (Garlikov, 1998; Rud, 1997).
However, the gap in the literature that now needs to be addressed deals with the power differential between
participants and whether or not this could influence the interlocutor's responses in a Socratic discourse. Is it possible
that Rud's criticism (even though he offers it as just an aside) of Socratic pedagogy is misguided, and assent to
propositions are the consequence of power dynamics rather than students being led to certain conclusions (Rud,
1997)? This essay will focus on these ideas, specifically exploring the nature of power in discourse as it relates to
Socratic questioning, and show that while the criticisms definitely have merit, they are not strong enough to
undermine the Socratic project.

There are two ways that power relations could impact a Socratic discourse, one obvious and one less obvious, if: 1)
the participants respond in a certain way because they seek something other than the truth, such as approval or a good
grade, and 2) the race, class, gender, ethnicity, sexual orientation, etc., of either the Socratic practitioner or of her
interlocutors play a role in the discourse, i.e., if arguments and counterexamples offered do not stand or fall on their
own merit, but because of an intrinsic quality of the utterer. Let us now examine these and see what role, if any, they
play in the successful practice of the Socratic pedagogy.

The relationship between knowledge and power in discourse has been extensively examined (Popkewitz & Brennan, 1997; Boileau, 2000). Often these criticisms focus on the more obvious abuses of power in discourse, such as individuals not being allowed into the discourse, or individuals who go into a discourse with certain assumptions about what someone can know based upon their sex or race.(Note 1) These are issues in any discourse, and the first point, while admittedly important, is structural and somewhat less interesting, and consequently will not be addressed here (i.e., in a classroom environment issues of self-selection of participants, or who physically gets to be in the classroom, is not immediately relevant to the ideas being examined here). The second issue does indeed impact Socratic discourse, and surprisingly little to no research has explained how power dynamics impact Socratic practitioners and their students (Boghossian, 2001). If it is the case that truth seeking educational communities cannot be established because of power disparities between students and teachers, then not just Socratic pedagogy, but the genuineness and authenticity of all dialogical pedagogies are called into question. If the problems posed by the Socratic teacher are met with responses that have some other intent rather than to get at the truth, then Socratic pedagogy cannot be said to be genuinely truth oriented because the participants did not yield to propositions on the basis of reason.

Race, class and gender

One of the presuppositions of the method is that what is at issue is the force of argument, not exogenous factors such as the race, gender or social class of the person who responds. But racism, sexism, and other isms do exist. And evidence shows that both student and teacher expectations are at least marginally determined by these unmitigated factors (Steele, 1998; Steele, 1999). These are, at least at the present time, tragic facts of life, and of course are not particular to Socratic pedagogy, but rather starting conditions that all educators encounter (Levin, 2001).(Note 2)

But is Socratic pedagogy more or less susceptible to issues of race and gender, and does this cripple, or at least negatively impact elenetic accomplishments? If, for example, a student's judgment as to the truth or falsity of a proposition is influenced by the white hair and upper-middle class mannerisms of the teacher, then both the Socratic process, and the conclusion it yields, could be suspect. In a classroom situation a Socratic practitioner could unconsciously discount a student's argument because they are, for example, of African descent.

But what is more likely, that an argument will be subconsciously discounted because of the race of the person who makes it, or that an argument, regardless of the race of the person who makes it, succeeds or fails *because of* the elenetic process? That is, in genuine Socratic practice arguments cannot be *de facto* rejected; they must be rejected because of a counterexample or by sheer force of argument. (Of course anyone can intentionally disregard statements by people of a certain race, and these more obvious and even more egregious instances are not at issue here because this has nothing to do with Socratic pedagogy and everything to do with blatant racism. What is at issue are people's voices being heard and their claims being answered, or not, because of who they are). Subconsciously or otherwise, of course the Socratic teacher could overlook, or give less attention to, someone's claims because of their race. One could, for example, disregard a devastating counterexample as irrelevant because they had the prejudice going into the discourse that people who are a particular race, gender or sexual orientation could never saying anything substantive. But this would not be Socratic; this would be a form of abuse that masquerades as Socratic and as such could be found in any pedagogical model. The claim here is that this is more and not less likely to be exposed in Socratic pedagogy due to the ability of rational participants to assent to true propositions; and this, in turn, is because of a rational process that removes much of the ambiguity and confusion from adjudicating claims.

The elenchus does not necessarily bring one's racial and gender assumptions to the surface, but it does force the participants to focus on the arguments and not the people who make the arguments. If there is ever a dispute, the *claim* is at issue and not the person. Because of this it is more likely that issues of race and gender will not play a role in the discourse, as opposed to other models where there is no process for the adjudication of claims. Therefore, while race and gender play a part in all dialogical contexts, they play less of a role in a Socratic discourse. As such racial and gender issues do not compromise the integrity of the Socratic method.

Power dynamics

The Socratic method centers on the notion that attaining the truth is possible through discourse (Vlastos, 1994). The idea behind this is that through argument, example and counterexample, rational participants will assent to true propositions. However, this is bundled with a number of presuppositions, such as the presupposition that participants enter into the discourse freely (as opposed to being forced to take a required class where the teacher's pedagogical model is Socratic), and that responses are being given because they are they are believed to be true (as opposed to being assented to because of convenience or because respondents will "get something" from their interlocutor).(Note 3) If it is indeed the case respondents will receive some tangible benefit, or at least perceive that they will, it stands to reason that they will provide answers that they believe the Socratic practitioner wants to hear. If they provide responses for any reason other than the belief that what they say is true, then the elenchus cannot achieve its epistemological ambitions. If this is the case then it is not a trick of inference, or a "twisting" of logic, but that the respondents want to give certain answers because of something *other than logic*, like approbation or even fear of looking stupid.

An important question then becomes whether it is the case that because of one's position as a teacher and the authority and power that come with that role, students in a Socratic classroom environment will assent to certain propositions

that they would not otherwise agree to if they were just with their peers? It is certainly possible, because of the student/teacher power dynamic, students would too easily permit the teacher to influence and even guide their responses. This has the obvious impact of subverting genuine educational discourse because differences in power between and among those engaged in conversation prohibit an honest exchange of ideas—and an honest exchange of ideas rests at the heart of the elenchus. For the elenchus to work, students need to agree or disagree with certain propositions because of their belief in their truth or falsity. So if it is the case that a proposition is offered not because it is viewed as being true, but for some other reason, then genuine discourse would seem to be inhibited. If students and teachers cannot have an authentic truth-seeking classroom, or even have a genuine discourse, then one of the principal goals of Socratic pedagogy—truth seeking—is seriously compromised.

Thus successful Socratic pedagogy will disabuse participants of more rigid notions of relations of power that are structurally embedded in traditional communicative contexts. Traditional power relations, specifically in a classroom setting, center on both the teacher's "power position" and her privileged access to the truth (Etzioni, 1975, p. 5). But paradoxically Socratic pedagogy confuses, and to an extent even inverts, traditional power relations. The Socratic practitioner is not claiming to have all the answers. She is, in a very real sense, deriving power from the declarations of her interlocutors (if there are no claims made the Socratic questioner has nothing to proceed from). When students participate in a Socratic discourse it is not immediately clear where the lines of power are. Truth is no longer the exclusive province of the teacher. Truth switches from people to propositions. In traditional discourses perceptions of truth are at least partially constructed by position, race, social and economic class, and even by aspects of appearance, like age or disability status. This reorientation of the power dynamic can be socially, intellectually, and even educationally disorienting.

Of course this does not negate the fact that participants in a Socratic classroom setting will respond in certain ways not because of the truth but because of a perceived benefit from a given response. It is not philosophical naïveté to claim that no matter what the reason is for one's responses, perceptions of reward may make students more easily led by the teacher, but it will not change either the truth of the matter or the defensibility of their claim. Perhaps this is best seen with a specific example from Garlikov's article:

> An example of the latter case was in a discussion of homosexuality in an "Ethics and Society" course where many students said that homosexuality was wrong because (the idea of) it was so disgusting. I asked them whether they thought that such disgust was a sufficient characteristic to make an action be immoral. They said it was. I asked them then to close their eyes and think about ... their parents having sex with each other. They all let out an even bigger groan of disgust, and said they found that idea really disgusting. So I asked whether they would have to conclude then that it was immoral for their parents ever to have (or to have had) sex with each other. They agreed it was not. Of course they then asked whether that meant I thought homosexuality was moral. My response was that whether it is or is not is simply unrelated to whether it is personally disgusting or not to anyone. I was not trying to argue in this particular case for or against the morality of homosexuality, but was merely trying to get them to see that finding an action disgusting did not justify their thinking it must be immoral just because of that (Garlikov, 1998).

In this outstanding example of the Socratic method, if students thought that their teacher did not like homosexuality, then they could easily have lied and given false statements. For example, anticipating where he was going, they could have responded that envisioning their parents having sex was not disgusting, but that it made them uncomfortable. This would still have left room for defending their claim that all things that are disgusting are immoral. But then Garlikov could have made further inquiries about other things that are disgusting, such as eating a plate of live insects, and shown that disgust is neither necessary nor sufficient to judge a thing as immoral. *In either case, no matter what their responses were, through successful elenetic inquiry a truth of the matter would have emerged.* Their claims would have withstood the elenchus, or not. The relationship between being disgusting and being immoral would have been established, or not.

So the question then becomes, how much, if at all, a student's giving a response that she think the teacher wants to hear is going to adversely affect the truth seeking conditions of the dialectic?(Note 4) My claim is that if the elenchus is successfully applied, power relations will still impact Socratic discourse, but not to such an extent as to make it an ineffective pedagogy. Not only is truth seeking not compromised, but also other virtues such as getting students to think critically and engage ideas remain unscathed. In our present example, to even think so far ahead in a discourse as to be able to anticipate where it is going requires a fairly high degree of cognitive ability.(Note 5) And if students are not capable of this, then the issue that they would give a response because of a teacher's sentiment, or because they want to "get something," are dulled. The idea of giving a response because of something presupposes that students know what that response is that they are supposed to give. Not only is it often unclear what response the teacher wants, but that does not guarantee that particular conclusions could be reached.

So then the issue becomes, what if students give responses not based upon the teacher's sentiment, but because they think that is the smartest response to give, and giving the smartest response means that they will get the best grade. (That is, the smartest response may not be one that a student believes accords with the truth, but the one that makes them look the most intelligent; so one's motivation would not be for the truth but to look intelligent.) Well, this still would not adversely impact the discourse to such an extent that its practice would be jeopardized. Giving the best response, or at least attempting to, would relegate the truth seeking status of the method to secondary or even tertiary significance, conveying primacy on the critical thinking aspect of the method. Depending upon the teacher's desires,

47

this could actually be beneficial.(Note 6) But this would only adversely affect (perhaps more by slowing down the discourse by taking more time to arrive at conclusions), and not endanger, the method's truth seeking orientation.

Conclusion

Garlikov addressed the first part of Rud's criticism about Socratic dialogue being leading. This work has addressed and extended his response to cover general notions of power, and shown how these could impact a Socratic discourse. Because of the proposition oriented nature of the elenchus, Socratic pedagogy is not merely an illusory exercise where participants acquiesce to notions of truth due to power differentials. But power relations certainly do play a role in all communicative contexts, and Socratic dialogue is no exception. What is an exception, however, is that the adverse effects of power are minimized, and the focus is shifted from people to propositions.

Notes

[1] For example, in the Symposium Socrates asks the women and the slaves to leave the room. Or more recently, feminist epistemology claims that the sex of the knower at least partially determines what is known, what can be known, and how it becomes known.

[2] The best way, if at all, these could be controlled for would be through blindly graded exams, probably utilizing a banking pedagogy where there are very specific right and wrong answers that need to be memorized and regurgitated (Frierc, 1970).

[3] Foucault would argue that one always gets something from being correct in every discourse, not just restricted academic discourses. Perhaps due to the limited context, it is more obvious what a student "gets" when he answers a question correctly. Where he stands in the power web becomes more visible. He gets a special relationship to the teacher. The teacher knows best and now he knows second best—and everyone knows that he know second best.

[4] What often happens in the classroom is that a good Socratic teacher is able to prevent students from correctly guessing what she wants to hear. This is because the Socratic teacher is inquiring into the reasoning behind a position—she examining whether or not it will stand up to scrutiny. Challenging a student's reasoning tends to make the student think that his conclusion is being challenged. Students very quickly learn that it is difficult to figure out the teacher's position, particularly when she challenges conclusions that are contradictory to each other, one of which is supposedly what the teacher believes. But if Socratic teachers are looking for sound arguments, and if the student is able to come up with a good argument, reason (and therefore the best method we have to search for truth by using evidence to make inferences and deductions) is served even if it also pleases the teacher. But the enterprise is so difficult in most complex situations that it is hard to imagine a student's coming up with a chain of reasoning that will withstand the teacher's scrutiny just because that student is trying to impress her or get her to like him by guessing. Guesses are not likely to do the job.

[5] In a personal correspondence Garlikov wrote, "even in the Socratic dialogues, as in classrooms, interlocutors give wrong answers that they try to support, which shows, I think, they are not just giving psychologically prompted answers, but answers that show they really think about the material—logically and conceptually."

[6] Though in my personal opinion, this would be a heartbreaking consequence of privileging intellectual qualities over a search for and love of the truth.

References

Boghossian, P. (2001). *The Socratic Method (or, Having a Right to Get Stoned)*. Paper presented at the *International History, Philosophy and Science Teaching* conference in Denver, Colorado, November 7-11, 2001.

Boileau, K. (2000). *Genuine Reciprocity and Group Authenticity*. Boston, MA: University of America Press.

Deluze, G. (1988). *Foucault*. Minneapolis: University of Minnesota Press.

Etzioni, A. (1975). *A Comparative Analysis of Complex Organizations*. New York, New York: The Free Press, Macmillian Publishing Co., Inc.

Freire, P. (1970). *Pedagogy of the Oppressed*. The continuum International Publishing Group: New York, New York.

Garlikov, R. (1998). Contributed Commentary on Volume 5 Number 20: Rud "The use and abuse of Socrates in present day teaching." *Educational Policy Analysis Archives*. Available http://epaa.asu.edu/epaa/v5n20c1.html

Levin, S. (2001). Social Psychological Evidence of Race and Racism. Chapter 3 in Compelling Interest: Examining the Evidence on Racial Dynamics in Higher Education. Report of the AERA (American Educational Research Association) Panel on Racial Dynamics in Colleges and Universities. http://www.aera.net/reports/dynampp.pdf

Popkewitz, T. & Brennan, M. (Eds.) (1997). *Foucault's Challenge: Discourse, Knowledge, and Power in Education*. Teachers College Press, Teachers College, Columbia University.

Rud, A. (1997). The use and abuse of Socrates in present day teaching. *Education Policy Analysis Archives*, 5(20). Available: http://epaa.asu.edu/epaa/v5n20.html.

Steele, C. (August, 1999). Thin Ice: "Stereotype threat" and Black College Students. *The Atlantic Online*. http://www.theatlantic.com/issues/99aug/9908stereotype.htm

Steele, C. (1998). Stereotyping and its threat are real. *American Psychologist, 53*, 680-681.

Vlastos, G. (1994). *Socratic Studies*. Cambridge: Cambridge University Press.

Zou, Y. and Trueba, E. (Eds.) (1998). *Ethnic Identity and Power: Cultural Contexts of Political Action in School and Society*. New York: State University of New York Press.

About the Author

Peter G. Boghossian
School of Education
Portland State University
Portland, OR 97207

E-mail: pete@boghossian.com
Homepage: http://www.boghossian.com

Peter Boghossian is currently writing his dissertation, in education, at Portland State University. He is an adjunct professor of philosophy at the University of Portland and Linfield College, teaches critical thinking in grades K-12 for Saturday Academy, teaches education classes at Portland State University, and is the assistant department chair for humanities at the University of Phoenix.

49

50

Education Policy Analysis Archives

Technology is Changing What's "Fair Use" in Teaching—Again

Linda Howe-Steiger
University of California, Berkeley

Brian C. Donohue
University of California, Berkeley

Abstract

The Doctrine of Fair Use was established by the courts to exempt certain activities such as teaching and research from the legal requirements of the copyright law. Before the 1976 Revision of the Copyright Act, only two cases were brought against teachers for copyright infringements. In both cases the teachers lost because their extensive copying was found to impact the copyright owner's market for legally published copies. Although the 1976 Act explicitly recognizes the existence of potentially Fair Uses, the act makes application of the principle highly situational. Classroom Guidelines attached to the Act make application even more murky and constrained. After 1976 photocopy technology and the advent of the coursepack began a trend towards circumscribing situations in which Fair Use may be applied. Potential impact on a new, lucrative market for sale of rights to copy portions of books and journals appears to dominate contemporary case law. Desktop publishing and Internet and web-based teaching, the authors believe, will further erode traditional applications of Fair Use for educational purposes. They argue that instructors and researchers should assume that there is no Fair Use on the Internet. Guidelines are provided for faculty and others considering dissemination of potentially copyrighted materials to students via digital technologies.

Ask any teacher in the United Stated whether or not it's "fair" to make free use of copyrighted materials in the classroom and his or her answer will most likely be, "Of course it is." Ask that same teacher why it is so obviously "fair" and you will probably get a blank look. Teachers just "know" that education has important social benefit and that they as teachers are exempt from usual legal obligations surrounding use of copyrighted materials. Or are they?

Introduction

Historical Perspective The Doctrine of Fair Use was conceived by the courts. It exempts certain categories of activity in some instances from the legal obligation to obtain permission from the author of a work before copying, performing, or displaying that work. Potentially exempt activities include teaching, research, scholarship, reporting, commentary, and even parody. The justification for the Fair Use exemption derives from the court's view that sometimes free and open discourse about ideas can be more of a stimulant to creation of new knowledge and new

creative works than protection of the author's ability to reap financial reward from his work.. Traditionally the use of excerpts from copyrighted materials for classroom teaching has been conceived as a Fair Use. In 1976 the U.S. Congress formally adopted the Doctrine of Fair Use into its revision of the Copyright Act.

That was twenty-five years ago. Since then technologies for reproducing, copying and displaying copyrighted materials have changed dramatically, and the locus for teaching activities has expanded beyond the classroom to include the airwaves (as in educational TV) and now the Internet. These changes have affected authors', teachers', and publishers' perspectives about is "fair" and what is not "fair." Today as the educational community moves rapidly towards web-based education and a growing emphasis on distance learning, we believe it is important to take a another look at Fair Use and its relationship to evolving instructional technologies, if only to protect schools, teachers, and course developers from unexpected legal challenges.

First, however, to understand Fair Use and its application to education, one must give up any idea that "fair use" was ever really about equity. It's not. Like copyright itself, the doctrine derives ultimately from Western concepts of individualism and principles of market- based capitalism. The identification of what is "fair" or "not fair" is deeply entwined with the nature and ownership of the technologies used to reproduce or distribute the works in question as well as to who stands to gain or loose from a particular type of use.

Constitutional Perspective

The basis for copyright is established by Article 1, Section 8.h of the United States Constitution, which states: "The Congress shall have power to promote the progress of science and useful arts by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." Copyright (along with patent, trade secret, and trademark) is the tool that implements the Constitutional purpose.Copyright is a legally enforceable intellectual property right that protects a financial incentive designed to encourage individuals to take the risk of creating new works or improving on previous ones. The productivity and innovation that lie at the heart of US economic success seems to provide clear testimony for the wisdom of the framers of the Constitution..

Educational Case Precedents

Before the Copyright Revision Act of 1976, the record of cases involving Fair Use of copyrighted material in the classroom is sparse. What exists draws on the general principle that copying for classroom use without obtaining permission from the author was not an infringement of copyright as long as there was no impact on the sale of published books or sheet music. There were only two cases concerning classroom copying by instructors before 1976. Both addressed the question of what portion of a copyrighted work could be freely copied and distributed to students under the Doctrine of Fair Use.

The earliest of these, Macmillan Co. v. King 223 Fed. 862 (D.C. Mass. 1914), was brought against a Harvard tutor who produced for his students very detailed outlines and summaries of an economics textbook published by Macmillan. Macmillan argued this was an infringement of copyright and negatively impacted their market. Although most students did purchase the classroom text, some did not, apparently relying solely on their tutor's materials. The court ruled against the tutor.

In the next case, Withol v. Crow 309 F.2d. 778th Cir. 1962, fifty years later, the court similarly ruled against a music teacher, who, short by forty-eight copies of a musical score for his students, decided to make copies rather than purchase additional scores from the publisher. These two cases represent essentially all of the case law defining the application of Fair Use to teaching activities until the 1976 Copyright Revision Act. Clearly it did seem at that point that if a teacher stayed safely below some upper limit of copying an entire work, then the teacher would not risk a complaint that her activity was not Fair Use.

The 1976 Copyright Law Revision

After 1976, however, Fair Use became more complicated for teachers. Even as the Congressional revision generally identified copying for educational purposes a potentially Fair Use, it laid a foundation for confusion by setting forth criteria to use in the determination of whether or not a specific instance of copying was actually a Fair Use. The analysis of Fair Use thus became highly situational. Two sections of the Act are directly relevant for this discussion.

Section 107 specifically states that making multiple copies of copyrighted materials for use in a classroom is not in itself an infringement of copyright. It then defines four factors that are to be used to analyze any specific situation—and so enters uncertainty:

> In determining whether the use made of a work in any particular case is a fair use the factors to be
> considered shall include:

> 1. the purpose and character of the use, including whether such use is of a commercial nature or

52

is for nonprofit education purposes;
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

In interpreting this section, the courts thus far have generally viewed the forth factor, potential market impact, as the most important factor, given that the commercial and financial monopoly is at the heart of the concept of copyright. However, the use of the word "includes" rather than "are" in this section opens the door to a suggestion that there could be other factors, perhaps more important, influencing the interpretation of a particular situation.

The second relevant portion of the Act is Section 110, paragraphs 1 and 2. This section establishes that Fair Use may apply to performance or display of copyrighted works during "face-to-face teaching activities" that are "a regular part of the systematic instructional activities of a government body or a nonprofit educational institution." The language here seems to focus on making a careful distinction between what might be construed as Fair Use for educational purposes and some other use that might be construed as either "entertainment" or a commercially motivated performance or display. The limitation of non-infringing performances to only those used in "face-to-face teaching activities" and the introduction of the vague concept that the non-infringing performance must be "integrated into a systematic course of instruction" further increase the complexity of applying this doctrine in the contemporary educational environment where digital instructional technologies allow teachers to download audio, video, graphics, text, photography, and radio and TV-like "webcasts" and "display" them to students inside or outside the traditional classroom via a class website.

The Classroom Guidelines

Incorporated by reference into the Act is a set of "Guidelines for Classroom Copying in Not-for-Profit Educational Institutions." This document provides detailed examples of how to implement an appropriate balance of private intellectual property rights of copyright owners (usually the large commercial publishers) and the public benefit that may result from unrestricted educational uses of copyrighted materials by teachers and students. The guidelines were developed by a diverse group of copyright stakeholders. Congress's purpose in incorporating them into the Act, according to the Congressional Record, was to demonstrate wide consensus on the application of "Fair Use" to educational practice.

We suggest that the Congress may have been deluding itself. In fact, the Congressional Record also reveals there was some real disagreement between those stakeholders interested in maintaining copyright free from any significant educational entitlements (that is, publishers and authors) and the academic community which views Fair Use for education purposes as an historic privilege. Educators on the committee also argued that teachers should be excused from obtaining permissions because the very process of obtaining and paying royalties to use materials would be an onerous duty and an inhibitor of academic freedom.

Indeed, the guidelines are very conservative and are increasingly difficult to apply in practice. Part of the reason they sound so silly today is that in being so specific they have simply not kept pace with changes in copying technologies (photocopying and computers). The guidelines set a very constrained standard for what may be construed as a "Fair Use" in an educational setting and may appear to contradict a more expansive interpretation of the language in the statute itself. For example, the guidelines define "brevity" as not more than 250 words of poetry, or 2500 words or less from a complete article, short story or essay, or 1000 words or 10 percent (whichever is less) from any prose work.. Copies of these brief excepts are non-infringing only if they are also "spontaneous" (i.e., according to the guidelines, a last-minute inspiration of the teacher) and one-time. According to authors of the guidelines, the basis for an exemption for such spontaneous one-time use of copyrighted works was simply to allow teachers enough time to obtain permission for the next classroom use (a process that in the mid-1970s took from four to eight weeks). Lost is the court's concept in the original doctrine of Fair Use that there was general social benefit in open discourse which itself was an encouragement of new ideas and innovation. Finally, the guidelines make clear that any post-class distributions of teaching materials to an interested but non-registered student could never be construed as Fair Use. (The guidelines may be viewed on-line at www.ucop.edu/ucophome/uwnes/copyrep.html, Appendix I.)

Why did the academic community sign-on to the guidelines if they were not practical? One answer may involve the technology for classroom copying in 1976. The mimeograph machine was cheap but it was also messy and irritating. The danger of teachers seriously impacting revenues from the sale of books and periodicals by making copies on the departmental mimeograph machines probably seemed fairly remote even to the teachers themselves. Indeed, in practice many teachers may have felt the constraints of the guidelines tight but livable given the situation. Anyway, it was standard teaching practice at that time, particularly in colleges and universities, was to send the students to the library reserved reading room.

Still, there were those who were wary. The American Association of University Professors and the Association of Law Schools joined in arguing that the guidelines failed to take into account the reality of how teachers actually used materials for teaching purposes (122 Cong. Rec., H 10, 880-81). They argued, more to the point, To protect themselves, many Universities today still disseminate the 1976 guidelines to their faculties, probably with tongue deeply embedded in cheek..

53

Emergence of Photocopy Technology

Shortly after 1976 came the photocopy revolution. With it libraries and instructors had a means to quickly, easily, and cheaply reproduce quantities of materials for research and teaching. Instead of gathering books and journals onto the shelves of the reserved book room where students lined up to read assignments from the one or two available copies, libraries and instructors could just hit the "number of copies" button on a big machine and in minutes have copies for even the largest class. The old departmental mimeograph machine went to salvage and a typical student excuse for not doing the reading disappeared. Instructors also felt themselves freer to pick and choose reading material for their students without being bound by selections in someone else's textbook.. Commercial copy centers sprang up around campuses, reserve reading rooms dimmed their lights, and the "coursepack" was born.

Meanwhile publishers and authors saw the photocopy machine as creating a whole new market for the sale of rights to reprint portions of books and articles from journals. Two cases established the rules for determining Fair Use in this new technical environment. The first was Basic Books v. Kinko's 758 F. Supp. 1522 (S.D.N.Y. 1991). In this case the publisher, Basic Books, challenged Kinko's failure to obtain permission from the copyright holders (usually the publishers) when reproducing coursepacks. Although Kinko's argued that permission was not needed because the coursepacks were for classroom use and hence were exempt under Fair Use, the court agreed with Basic Books. Kinko's had unfortunately attracted the attention of the publishing world by advertising its ability to produce quick turn around of copies because it did not have to take the time to obtain permission from publishers.

An analysis of the Kinko's case emphasizes the commercial nature of copyright law and demonstrates how the Doctrine of Fair Use may be modified when a new technology creates opportunities for business. The court held that Kinko's failure to obtain permission had negatively impacted the market in permission or licensing fees. Kinko's had "extinguished" a financial reward to the copyright holder (the publisher), which was precisely the reward that copyright was designed to protect. Today most copy shops rigorously refuse to reproduce coursepacks unless permission to reprint has been granted in writing. Any permission fees charged by the copyright owner are then passed on to the students. The key to understanding the Kinko's case is to see that the court did not really address the Doctrine of Fair Use as it applies to actions by instructors and students. Instead, with typical narrowness, the court focused the discussion entirely on the two businesses in the middle of that educational relationship. Since the copy shop had a commercial interest in the coursepack, the copy shop could not view its own reproduction as "Fair Use," despite the end use of its product for classroom teaching purposes.

The principle laid down in Basic Books v. Kinko's was repeated, clarified, and perhaps strengthened in a second case on the same issue: Princeton University Press v. Michigan Document Services 99F.3d. 1381 (6th Cir. 1996). In this case, Michigan Document was one among several copy shops operating near campus. Its owner deliberately set out to test the ruling in the 1991 Basic Books case. Michigan Document therefore, not only did not obtain permission to reproduce materials for coursepacks, but also advertised this action, passed the savings incurred on to students, and used this reduction in price to undercut competitors. It is not surprising Michigan Document drew the attention of the publishers, and a negative ruling from the court.

Again, however, the key to understanding the court's ruling is the commercial exploitation of copyright. In brief, the court asked: who is making money from the copying and who is losing money? The court reaffirmed that if the copier (Michigan Document) makes money from the copying, then the copying could not be construed to be Fair Use, even though the reproduction had an ultimate educational purpose. The court amplified that the Fair Use exemption is not a blanket exemption and that when litigants have commercial interests, the burden of proof that a copying situation is "fair" lies with the copier—in this case the copy shop. Since ultimately the copy shop is in the business of commerce, that is in making money not in teaching, the Doctrine of Fair Use does not apply. The court also again noted the existence of a lucrative new market in permissions created by photocopy technology. Teachers who believe they can "get around" the commercial aspect of the copyshop decisions by making their own copies in the library or on their own office copy machines might do well to take another look at the story of the Harvard tutor back in 1914.

The court has never actually tested the legitimacy of the coursepack itself, just of its reproduction by a commercial business. Essentially the coursepack is a unique collection of materials assembled by an instructor for a particular class that may be delivered one or more times. Components of the "pack" are quite long—whole chapters, articles, essays, or stories—far exceeding the so called safe harbor standards presented in the classroom guidelines. Under the 1976 Act such a collection itself is a copyrightable work, referred to as a "collective work..."

Digital Publishing and the Internet

The digital technology for both desk-top publishing and distance learning, including webcasting, class web-sites, e-learning, and in-class real-time Internet access, is here now here. With it has come a quantum leap in the murkiness of applying the Doctrine of Fair Use for education. Not only does the approach suggested by the 1976 Act seem outdated, but also Congress's recent effort to update copyright for the computer age—the 1998 Digital Millennium Copyright Act—deliberately sidesteps many of the toughest issues for educators by declaring them simply "unsettled." (See the "Report on Copyright and Distance Digital Education," May 1999, US Copyright Office available at Error! Bookmark not defined..)

The first "unsettled"—if not "unsettling"—issue concerns the definition "copy" on the Internet. When an on-line instructor assigns a student something to read something and the student "retrieves" that something from a digital file on a server connected to the Internet into her desktop computer, is that a copy? According to the Copyright Act, to be a "copy" the reproduced work needs only to be fixed in a tangible medium of expression. In Advanced Computer Services v. MAI Systems Corp., 845 F. Suppl. 356 (E.D. Va. 1994), the court ruled "the representation created in the RAM 'is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration'." In other words, "Yes," the student is reading a "copy." But is this copy a legally reproduced copy or a copyright infringement?

Although the focus of Advanced Computer Services was software, the principle laid down by that decision would seem to be applicable to any work, including articles and Power Point presentations, or video clips, viewed or read via the Internet. If that is so, it raises the spectre of every single viewing of copyrighted materials over the Internet involving either a potential copyright infringement or a royalty payment. Will everything be pay-per-view? This interpretation seems overly restrictive, undercutting the balancing of interests that copyright law attempts to achieve between public and private benefits. While authors do need protection of their ability to reap benefits both in dollars and reputation, the ultimate goal of copyright—the encouragement and advancement of knowledge and creativity—would not be served by skewing everything in the direction of the author's monopoly of copyright.

Indeed, the court in Religious Technology Center v. Netcom On-Line Communications Services. Inc., 907 F. Supp. 1361 (N.D. Cal. 1995) took a more reasonable approach when it characterized what happens when a user browses through materials on the Internet as "the functional equivalent reading"—not copying. Judge Whyte viewed this such reproduction of a file on the computer screen as simply a necessity for humans to perceive the information—rather like stand-up reading in a book store or borrowing a journal from the library. We hope, in the end, that Judge Whyte's reasonable view will prevail. Anti-copying technology certainly exists to disable a reader's ability to print, attach, or email an article over the Internet—activities that may be closer to a traditional understanding of copying. However, such distinctions are by no means without uncertainty in the evolving legal environment.

How, then, will the Fair Use Doctrine be applied to education and teaching in this environment? Unfortunately, current trends appear to be towards protecting commercial interests rather than protecting the public's access to knowledge or learning. We note that the instructor and the learner in any on-line situation are often separated by one or many commercial interests—for example an Internet service providers (ISPs) or an e-business providing course management or courseware to an instructor. One would do well to recall the Kinkos and the Michigan Document decisions.

In addition, technology may help resolve the complaint made in the minority report on the classroom guidelines that the task of obtaining permission to use copyrighted materials is onerous and time consuming. This, at any rate, is one possible implication to be derived from the decision in American Geophysical Union v. Texaco, 802 F. Supp. 1 (S.D.N.Y. 1992). In this case the court ruled against a Texaco researcher who, for his own convenience, regularly made photocopies of articles from journals which he kept in his office for future reference. Though such copying is common practice among both researchers and teachers, whether working in commercial or noncommercial context, the court accepted the copyright owner's argument that permissions were easy to get because of new Internet-based Copyright Clearinghouse technology and then went on to reason that since the ultimate purpose of the researcher's activity was profit for his employer (Texaco) that the private monopoly interest of the authors should prevail.

Somewhere the larger issue that led to the establishment of the Doctrine of Fair Use in the first place seems to have gotten lost in excitement over how easy the new technology can make the payments for copying rights flow. We as a society need to stop a moment and review what earlier courts had to say about the importance of support for the individual teacher or researcher who is exploring ideas and creating knowledge for the next generation and the general benefit of society.

Conclusion

There's No Fair Use on the Internet

The current use of Internet technology to support teaching brings new commercial players into the communications continuum, separating teacher and student in the so-called distance learning environment. Will this situation permanently eliminate the Fair Use exception for digital teaching? In all of the cases cited above (and there are no cases on the "other" side)—whether it is a commercial copy shop making coursepacks, an instructor copying sheet music for students, or an engineer copying articles for his research files—when there are royalties to be collected and any potential commercial interest in the vicinity that might be viewed as either funneling off or diluting a potential profit for the copyright owner, then Fair Use copying may not apply. Even the possibility of diluting some future profit may be deemed sufficient to establish an market impact.

In short, it is easy to conclude that there is no Fair Use on the Internet because the Internet per se has become a commercial, for-profit business. To those who would say that information on the Internet is in most cases free and not commercially motivated, we reply that the goal of Internet companies and those who support their development is ultimately to make money—if they don't, they'll go out of business. The collapse of the dot com bubble makes that

truth almost self-evident. Therefore, to help educators who are creating courses on the Internet (which are themselves copyrighted works), we have constructed some guidelines:

1. Know what's copyrighted and who holds that copyright. Before 1978 every copyrighted work had to carry a copyright notice. If it didn't, it wasn't copyrighted. So look. After 1978 the explicit copyright notice was no longer required for establishing legal ownership. Any fixed expression is copyrighted under common law from the moment of its creation, whether or not it is formally published or registered with the Copyright Office. This includes student work, institutional reports, neighborhood newsletters, and office memos. If something published since 1978 does not explicitly state that it is in the public domain and may be used and copied freely, it should be considered copyrighted material.
2. Don't assume something is in the "public domain" just because it's a government document. Everything published by a government agency or funded with public dollars is NOT necessarily in the public domain. Seek out a copyright notice or other notice regarding rights to reprint or post on the Internet. If you cannot find one, ask the author or publisher. Many non-profit organizations simply want you to send them some notification that you are using their materials for bragging purposes. Asking for copyright permission does not necessarily mean you will have to pay a fee.
3. Take personal responsibility for obtaining permission before using any copyrighted work. Should a suit be brought for copyright infringement, you personally will be liable. Don't assume that the Fair Use exemption will apply because your use has some educational purpose..
4. Never post copyrighted material (see rules 1-3) on the Internet assuming that you are exempt from obtaining permission by the Fair Use Doctrine because you are posting to a class web site or because you are an instructor or researcher making educational materials to anyone who's interested. There may be a commercial interest somewhere you will tread on. Explicitly seek and receive permission from the copyright owner. An alternative is to provide students with references for the library or the bookstore or point them towards the author's own on-line version or to a legitimate digital library managed by someone else who is presumably posting only legal copies.
5. Do not distribute copyrighted course materials that you formerly distributed via coursepacks over the web (even if you formerly obtained permission for the coursepack copy) unless you restrict access to materials to just the registered students in your class (i.e. a password protected class site) or have received specific permission from the copyright owner to make copies available on the web.
6. Always give full, standard bibliographic citation on the digital copy itself, including a statement that permission to reprint for use by your on-line students has been granted.
7. Allow yourself plenty of time to identify copyright holders and receive their permission to publish via the Internet. Despite the availability of an evolving digital clearinghouse technology, common sense suggests that everything will not be in the database and you are ultimately responsible.

About the Authors

Linda Howe-Steiger, Ph.D., AICP
Institute of Transportation Studies
Technology Transfer Program
University of California Berkeley
1355 S. 46th Street, Building 452
Richmond, CA 94804
Tel: 510-231-5678
Fax: 510-231-9591
Email: lkhs@uclink.berkeley.edu

Linda Howe-Steiger is Director of the Technology Transfer Program for the Institute of Transportation Studies at the University of California Berkeley. She holds a BA in English from Bryn Mawr College, an MA from the University of Chicago, a Ph.D. from the University of Pennsylvania, and Master of City and Regional Planning (MCRP) from Rutgers University. She has taught graduate and undergraduate courses at Rutgers is developing a series of on-line tutorials for transportation professions as part of the California Learn-Net project. She is a member of the American Institute of Certified Planners.

Brian C. Donohue, J.D., M.B.A.
Business Contracts Administrator
University of California Berkeley
6701 San Pablo Avenue, Suite 218
Berkeley, CA 94720
Tel: 510-642-3128
Fax: 510-642-8604
Email:donohue@uclink4.berkeley.edu

Brian C. Donohue holds a BA from Fordham University, an MBA in Information Technology from George Washington University, and a JD (Juris Doctorate). He received a Certificate of Advanced Trail Advocacy from Hastings College of Law and has taught "Intellectual Property Law and the Internet" for the University of California Berkeley Extension for six years. As Berkeley's Business Contracts Administrator he has executed on behalf of the Regents of the University over 100 Internet agreements.

Education Policy Analysis Archives

The Power-discourse Relationship in a Croatian Higher Education Setting[1]

Renata Fox
University of Zagreb
Croatia

John Fox
University of Zagreb
Croatia

Abstract

Croatian higher education system's public space is researched through a critical analysis of a Croatian faculty's discourse. Representing a typical faculty social situation, two council meetings— recorded in minutes—are critiqued. Both meetings' minutes provide evidence of discourse strategies of deception used by faculty power holders to create an illusion of consent. We attribute the success of the deception to council members' ideas about the Faculty's groups/individuals, relations and issues related to the Faculty's hierarchy, their rank within that hierarchy, and their position within the Faculty's social network. To support our argument, we explore how the Faculty power holders' discourse is built on a power/ideology/language formation. We conclude that, failing to critique the faculty's discourse, council members neglected their historical task of paving the way to democracy.

Introduction

According to research on the quality of teaching in Croatia's tertiary education conducted by Ledi et al (1998), the position of education and the teaching profession in Croatia is critical (cf. also Marinkovi Škomrlj 2000). Croatian students demand changes, but know that they do not have the power to initiate them: their role is compulsorily passivised (Ledi et al 1998:633). Students' complaints range from a general lack of communication (mainly due to teachers' inaccessibility and/or unwillingness) to fear (created through teachers' threats /ibid. pp. 630-632/).

Tertiary education in Croatia is financed by the Ministry of Science and Technology (MOST) and regulated by the Law of Higher Education and the Law of Scientific and Research Activity, both passed in 1994, and both marked by numerous "deficiencies, contradictions, lack of clarity and gaps" (Dika 1998:V). In 1998, MOST published a "Blue Book," offering an interpretation of the laws. Responsibility for a Faculty's finances is held by the dean who has the

authority to allocate funds at his/her discretion. In Croatia's deteriorating economy with growing unemployment—presently about 25% (out of which only 20% are entitled to some social benefit /Gali 2000:3/)—and low (teachers') salaries (Matkovi 2000:5), control over finances means power. For example, 10 per diems for travelling abroad approximately equals a teacher's monthly salary. There is a fierce battle going on out there between classes, groups, alliances and individuals: its purpose is to make or break relations of domination (Fairclough 1999a). Whereas the West clearly understands what democracy means, "postcommunist nations do not" (Savitt 1995:17). For them the emerging "new order" is only vaguely defined and understood. The scene, to quote Josip upanov, Croatia's leading sociologist, is a "unique combination of incompetency and corruption" (Srdo 2000:5), dominated by the need for power and money (Fox 2000; Fox & Fox 2001).

This article researches the Croatian higher education system's public space through a critical analysis of a Croatian faculty's discourse. For the purpose of this analysis discourse is understood as a collection of interconnected texts, i.e. communicative events wherein social, cognitive and linguistic actions converge (Beaugrande 1997:10).

Representing a typical faculty social situation, two council meetings—recorded in minutes—are critiqued. Each meeting minutes is evidence of a discourse strategy (the how and what to say) of an act of deception—an attempt to move "someone's thinking in a wrong direction" (Ng & Bradac 1993:118), where "wrong" means away from the speaker's real intentions or feeling—used by Faculty power holders to create an illusion of consent.

As we shall show, the acts of deception feature knowledge and opinions about teachers' (actual) selves, other council members (e.g. students), goals of interaction and important social dimensions of the council meeting itself. Critiquing the council meeting minutes, we shall demonstrate how "the semiotic (language, discourse in the abstract sense, text) figures as an element of the social" (Fairclough 2000:186). Put another way, we shall use the council meetings to systematically explore the "opaque relationships of causality and determination" (Fairclough 1999a:132-133) between (a) the council members' discursive practices and (b) wider social and cultural structures, relations and processes in the Croatian higher education system and Croatia itself. Finally, we look for a solution.

2.0 Council Meetings

The following two cases refer to council meetings held at the Faculty for Ore Searching, Blue River University, Croatia (henceforth Faculty).

Case 1: A dean election campaign

Setting: The Faculty for Ore Searching suffers from "massivity": over 2000 students in a comparatively small and inadequately equipped building (7 classrooms, 50 PCs). There is a pronounced tendency of increasing the total number of students through enrolling part-time students, paying students and opening dislocated departments. Due to sheer numbers, pedagogical standards defined by the Ministry of Science and Technology (MOST)—for example, the maximum permitted group size—are not respected. Students feel cheated and, consequently, are dissatisfied.

The winning candidate in a dean election campaign used the slogan "A chair for every student." He expanded the proposition in his program as follows: "From now the number of students to be enrolled will be limited and a new improved organisation of teaching processes will put a stop to overcrowded classrooms and lecture theatres."

He never kept his promise: In fact, the total number of students was increased. The dean's breech of electoral promise invited a reaction on two levels:

1.   Teaching and administrative staff kept pretending everything was fine. If, for example, asked how (academically, administratively, physically) this enormous number of students should be dealt with, the vice-dean for teaching would say with a saintly smile: "one day it will be better." Any reference to the mass of students was considered an attack on the Faculty. Teachers who, for whatever reason, had to refer to the issue at the Faculty's council meeting, would ritually start with: "Please do not misunderstand me. I would be the last person to criticise this house, but we seem to have a problem with numbers ...." As a rule, any reference to "massivity" is omitted from council meeting minutes.

2.   Students felt let down and objected—via their representative—to the situation at the council meeting. The students' representative quoted the election slogan *A chair for every student* and openly stated that the promise had not been kept. His objection was

---

disregarded; what is more, he was verbally attacked by his colleague, another student representative, during the meeting. His discussion was not mentioned in the meeting minutes. Following the meeting, he was advised not to "create any more problems." When informed about their representative's failure, students reacted with a graffito scribbled all over the Faculty: *Every student gets a chair, but not at this faculty*. The issue was never referred to again. It was omitted from meeting minutes. The student representative's objection to this omission, raised at the subsequent council meeting, was also omitted from the meeting minutes.

2.1 Case Interpretation

The act of deception here is built upon the metaphor *chair*, a shorthand for the dean's explicit promise stated in his program:

"From now on the number of students to be enrolled will be limited and a new, improved organisation of teaching processes will put a stop to overcrowded classrooms and lecture theatres."

The success of the metaphor *chair* can be explained through a 3 phase procedure: identification, interpretation and acceptance. The procedure is founded on Kintsch's (1989:185-209) construction-integration model of discourse processing. First, all possible meanings (not just the relevant ones) of a word are initially activated—the identification stage. Following this, non-relevant meanings are soon deactivated—the interpretation phase, while the relevant ones raise their activation and spread into the rest of the discourse—the acceptance phase. Operations within the model make it probable that out of the whole meaning potential of a word, the intended (ascribed) meaning is highlighted.

Related to *chair*, then, identification is the interactants' (dean's, teachers', students') awareness of the existence of the metaphor, enabled through shared social knowledge. In the interpretation phase, non-relevant meanings (e.g., building a new lecture theatre as a possible solution to the space shortage) are deactivated. Interpretation is a result of the interactants' agreement on the ascribed meaning of the metaphor (reducing the number of students). The power of *chair* lay in its transparency and conventionality which enabled it to become a part of the electoral body's social cognition. Captured by their wishes with which the metaphor perfectly harmonised, the voters oversaw its arbitrariness and eagerly consumed the offered virtual reality. Their inability (or unwillingness) for reflective analysis of the Faculty's discourse prevented them from seeing *A chair for every student* as an act of deception. Finally, in the acceptance phase, the ascribed meaning of *chair* raised its activation and spread into the rest of the Faculty's discourse. Acceptance confirmed the reality of the metaphor. For all those teachers (and students) who saw reality as defined by *chair* (cf. Lakoff & Johnson 1980: 156-158), the metaphor was true.

Case 2: Access to information

Setting: The dean is perceived by both Faculty staff and students as an autocratic manager who relies on coercive/reward power and forces compliance. He gives selected Faculty council members (both teachers and students) privileges (e.g., fees, grants) which they—in order to maintain social relationship—have to honour through obedience.

During his first year of office the dean assumed total control of the institution. All information about national/international conferences/events is collected in his office. The dean personally decides on who gets what information. According to him, allocation of information is in the best interests of the Faculty: teaching staff receive information related to their area of teaching and scientific affiliations. Every now and again, however, teachers do complain of not having been informed of a particular event (Note 2). A case in point is the Croatian association of Ore Engineers' annual meeting held in hotel *Glitzy*, Zagreb, which was attended by Faculty management. Surprisingly, teachers actually living in Zagreb were not informed. Another similar example is the International jubilee congress *Ore Searching and the Environment* held in China, which the dean attended. Information about the congress was never passed on to teaching staff. Again, criteria of access to information were questioned by teaching staff. It was pointed out that environmental issues are covered by three Faculty teachers in their graduate and postgraduate courses, who have published in this field both in Croatia and abroad (which the dean has not).

1st Council meeting

59

A member of the teaching staff, JJ, raised the question of access to information. He stated that not all teachers were being informed about national/international conferences, and that criteria for disseminating information were not transparent. He suggested, therefore, a central information portfolio, which would contain all conference-related information received by the Faculty, to be kept in the library, available to all staff.

The dean objected, claiming that conference information—depending on the conference topic—was allocated to relevant Faculty departments, and that the system worked just fine.

JJ pointed out that what he was suggesting was total communication, i.e., all research-related information available to all teachers, which, in his opinion, would be in the best interest of the organisation.

The dean called over three teachers (heads of departments for economics, management and foreign languages). They confirmed that the system worked just fine and there was no need to change it. Students did not participate in the discussion (Note 3).

2nd Council Meeting

JJ objected to the minutes of the 1st meeting. (Council meeting minutes are circulated to council members almost four weeks after the meeting, usually less than a week before the next meeting.) There was no mention of his discussion. He repeated his discussion and requested that it be included in the minutes. The recording secretary declared she would include it in the minutes.

3rd Council Meeting

The minutes of the 2nd meeting noted JJ's discussion as follows: "JJ objected to the minutes of the 1st meeting, as his discussion was not correctly interpreted." Again, JJ objected to this, stating that he had originally objected to the minutes of the 1st meeting because his discussion was *omitted*, not *wrongly interpreted*. The recording secretary consulted her notebook and said: "That's correct. This is my mistake. It will be recorded in the minutes."

4th Council Meeting

There was no mention of anything at all related either to JJ's discussion or secretary's admission of error in the minutes of the 3rd meeting.

2.2 Case Interpretation

The act of deception here is built on a continuum of misrepresentations (cf. Metts, 1989) of JJ's suggestion, ranging from omission of relevant information in meeting minutes to falsification—contradiction of truthful information. Omission is marked by silence; falsification by a particular lexical choice—the use of *interpreted* rather than *omitted*.

It would be a mistake to define the omission of information—in effect silence—negatively, i.e. as a mere absence of speech. All omissions in the meeting minutes had propositional content, which made them equally important to any formational unit of linguistic production and any element of discourse (Fox 2001:23). Both omissions (in the 1st and 3rd meeting minutes respectively) were made to produce an impression of consent, aimed at controlling public space.

Whereas the verb *interpret* represents a cognitive endocentric process featuring "inner" events and is associated more to mental or data-based activities, *omit* represents an exocentric process, featuring "outer" events and is associated more to behavioral or material-based activities (cf. Beaugrande 1997:208-213). *Interpreting* entails effort and attention on behalf of an agent, it involves a subject (someone who interprets). It is a process open to the agent's control, who can either initiate it or refrain from it. *Omitting* entails a simple non-effortful static quality. It emphasises the process in its own right. Whereas a process of *interpretation* has no direct target, *omission* tends to have a direct target, an affected entity expressed as an object, for example JJ's suggestions.

Use of the word *interpreted* instead of *omitted* in the minutes created inferences crucial to the success of the act of deception. The dean was portrayed as intensely cognitively involved in the Faculty's discourse. At the same time, he was removed from the discourse as the sole agent who had no direct target. The Faculty's discourse became a shared one with no affected entity. Briefly, skillful formulation of meeting minutes created a virtual reality where JJ's objection was lost, and, consequently, the dean's accountability to the objection. A potentially discreditable situation

was turned into a creditable one.

3.0 Explaining the success of deception

If, as our above anaylsis suggests, we accept that (1) faculty power holders committed acts of deception, and (2) assume that Faculty council members were cognizant of the deceptions, then, council members' absence of participation and criticism—expressed through their persistent silence—was, in effect, a vote of acceptance. Why did council members accept the deceptions?

We argue that their acceptance was a result of their social cognition of the Faculty's discourse: dominating social norms and rules which were simply translated into "specific constrains of discourse" (see van Dijk 1996:167). Social cognition is a socially shared system of social representations which may be "conceptualized as hierarchcal networks organized by a limited set of relevant node-categories. Social representation of groups, for instance, may feature nodes such as Appearance, Origin, Socioeconomic goals, Cultural dimension and Personality. These categories organize the propositional contents of social representations, which not only embody shared social knowledge, but also evaluative information, such as general opinions about other people as group members' Briefly, social representations include "socially shared cognitive representations, about social phenomena, including social groups, social relationships, or social issues or problems" (van Dijk 1996:166).

Social cognition includes too what van Dijk has called individual "(situation) models", i.e. cognitive representations of personal experience and interpretation, which include personal knowledge and opinions of other persons, of specific events and actions (van Dijk 1996:166). Models are the cognitive counterparts of situations.

3.1 Routine overlearned reaction of compliance

One social representation which perhaps explains council members' reluctance to criticise during the council meetings was a "routine, overlearned reaction" of compliance (Folkes 1985:133), subject to their involvement in the message (the lower the involvement, the less critical the information processing), their preparedness for the content of the message, and on their need for cognition (Petty and Cacioppo 1986). In a world where information production is much faster than information consumption, an individual's attention is inevitably diverted from message content to message form (Cialdini 1984; Redfern 1989). If the form of a message appears appropriate, the audience will be less inclined to question its content (Ng & Bradac 1993:133-134). The Faculty's council meeting minutes are excellent examples of *appropriate* "orderly interaction" (cf. Fairclough 1999a:28): an interaction which makes most participants feel (or pretend to feel) that things are "as they should be." A perception of orderliness was created by preplanned turn taking (during the meeting), which eliminated the need for a wider discussion. Individual power holders' turns (dean's, vice-deans', secretary's, heads of departments', "initiated" students' representatives") were fitted harmoniously together. Besides two individual challengers, nobody objected to either properties or effects of the discourse.

Both Case 1 (except for the student representative) and Case 2 (except for JJ) are marked by council members' low involvement and low need for cognition. In case 1 there was a manifest difference between teachers' and students' attitude towards the message (*A chair for every student*). Accepting the message at its face value, the teachers demonstrated a low involvement and low need for cognition. As the message was announced well in advance—copies of the candidates' applications for the dean's office were submitted to all members of the council two weeks before the council meeting—unpreparedness for the message is not an explanation (or excuse). Repeatedly questioning the validity of the message, students, on the other hand, manifested a high degree of criticism, high involvement and a high need for cognition. Their individual (situation) model was different to that of their teachers'.

Similar to Case 1, council members (teaching staff) in case 2 again demonstrated low involvement manifested in the absence of discussion (out of 36 teachers present at the council meeting, nobody expressed an opinion). As the agenda for the council meeting was distributed to council members almost a week in advance, unpreparedness, for the message, as in Case 1, cannot serve as an explanation for lack of response. Council members showed too a very low need for cognition. They uncritically processed information, and refused to accept the idea of total communication, which in itself aimed for an increased level of cognition. They rejected the idea of total communication not only within a particular social situation (council meeting), but also on a metalevel. Although aware of the fact that total communication would have been in their interest, students refrained from comments. Their low degree of criticism was a clear sign of low involvement and a low need for cognition. In contrast to Case 1, the students' behavior here was similar to that of their teachers.

To sum up, both case 1 and case 2 council meeting minutes show council members' overlearned reactance of compliance manifested in their low degree of criticism, low involvement in the message and low need for cognition. Individual variations (in Case 1 a students' representative, in Case 2 JJ) were a result of individual cognitive representations which were at variance from the rest of the council. We argue, however, that what really influenced council members' behaviour are social representations primarily related to the Faculty's hierarchy, their rank (title) within that hierarchy, and their position within the Faculty's social network (willingness to participate in relationships of authority, reciprocation and ingratiation). We suggest that it was the dominating group's power/ideology/language formation which influenced council members' social cognition, which, in turn, enabled the power holders to

61

successfully use the strategies of deception.

3.2 Power

The dean's authority gave him the legitimate right of function "to give orders and expect to be obeyed" (Smith & Vigor 1991:21). Creating obligations through giving selected members of staff extra assignments cum fees, which they had to honour through obedience to maintain social relationships, he was able to influence their behaviour through reciprocity. By enhancing the rewarding aspects of cooperation through pointing out similarities between himself and staff members, through self-deprecation and other-enhancement (Dickson et al 1993:1155), he was able to create ingratiation. For example, in Case 2, the secretary took the blame for incorrect minutes to save her boss from public embarrassment. Acting in line with rules of a reciprocity network, she symbolically reinforced her subordination, and made the dean's face "shine more clearly" (Jackall 1996:99). In return, she could hope for certain perquisites, such as protection for mistakes made, fees for "extra-assignments," benevolent treatment etc., as such is "the business of preservation of privilege" (Lomnitz 1977:206).

3.2 Ideology

Aware of changing alliances and balances, the dean knew that his power depended not only on his hierarchical position, but also on his ability to naturalise (Fairclough 1999a:27-35) his discourse through ideology: the greater the level of naturalisation, the more difficult it is to recognise discourse as an ideological representation of reality. For the dean, ideology became essential to produce and reproduce relations of power and domination (Fairclough 1995:14). He was able to use ideology "in the service of power" (Thompson 1990) simply because he had the social power to make his discourse seem as non-ideological "common sense", opaque and accepted as the Faculty's discursive norm.

Carrying the Faculty's ideology, the dean's discourse had to be perfectly naturalised. The instrument of naturalisation in Case 1 was a metaphor (*chair*) and in Case 2 lexical choice (*interpreted* instead of *omitted*). Naturalisation enabled the Faculty's ideology *The institution will serve personal interests of the power holders* to be turned into an opaque *Individual interests* (meaning, in fact, the interests of the dean's opponents) *must be sacrificed for a common goal*. As any other ideology, the Faculty's ideology was hidden, rendering the dean's power untraceable, and thus establishing an ideal frame for deception.

3.3 Power hiding

As argued by Ng and Bradac (1993:191), power hiding is the most subtle and complex of all power—language relationship models. The power-hiding effect of the dean's discourse is seen in contradictions between publicly declared intentions and actions. Whereas both the electoral slogan and meeting minutes claimed the Faculty's primary objective to be quality, in reality, promises were broken and channels of communication closed. Both Case 1 and Case 2 show a combination of, predominantly individual, *power to* and, predominantly collective, *power over* (cf. Ng and Bradac 1999:3). *Power to* is visible in the dean's efforts to achieve his personal goals and hinder others' achievements of goals (through total control of the Faculty and personalised criteria). *Power over* is visible in the Faculty's social network through which the dean maintains relations of authority, reciprocity and ingratiation (e.g. the secretary taking the blame for the dean). Both cases confirm Foucault's (1980) great truism that the residence of power is neither an individual nor an institution, but a network of human relations. Is there a solution?

4.0 Towards a Solution

Through control of knowledge and a strategic use of the Faculty's discourse, the dominating group is able to manufacture consent, essential for power/ideology reinforcement. The relationship between power and discourse at the Faculty has simply become a question of democracy, and, as Fairclough (1999a:221) emphasised, "those affected need to take it on board as a political issue."

Both teachers and students are in a need of social emancipation which, inevitably, is about tangible matters, such as the right to work, access to resources and distribution of wealth (cf. Fairclogh 1999b:233-234). With time, the dominated group's (most teachers, students) forced compliance should create a resentment which, if turned into an organised force and used as an impetus for a power struggle, could become an instrument of resistance. The rebellion of the oppressed will of course be resisted by the present power holders. As Faculty management tends to rely on coercive power, the first prerequisite of resistance will be courage (cf. Kreitner & Kinicki 1992). Only if enough audacity and persistence for a prolonged personal struggle is accumulated by the oppressed, can there be a "rising of consciousness" (ibid. p. 234), which in turn will empower the oppressed to engage in a struggle towards emancipation.

In their resistance, the Faculty's oppressed will need a leader, a person who will function as a "catalyst" (Fairclough 1999b:234). It is generally believed that the catalyst should possess two qualities: (1) some theoretical knowledge to be able to assume the role of a coach, and (2) the experience of the oppressed in order to gain trust and be accepted by the group. In real life, however, the catalyst is often a dissident from a dominating group who, anticipating change in the balance of power, somersaults into the new role. At any rate, it is through a catalyst's assistance that the Faculty's

oppressed will start learning how to deal with discourse-related issues of power.

The first sign of changing social representations and individual situation models of the Faculty's council members will be questions related to ideology and discourse:

- What is the Faculty's ideology?
- What is the relationship between the Faculty's ideology and the dominant discourse?
- How transparent is the Faculty's dominant discourse?

Learning about the effects of ideology upon discourse, and, in turn, of discourse upon ideology should help the dominated group to denaturalise the Faculty's ideology, i.e. recognize it as such. The bond between social determinations (a struggle for power maintenance) and discourse effects (naturalisation of ideology), previously opaque to many participants, will become clearer. Increased discourse awareness will provide the oppressed with the means to "challenge, contradict and assert" in an environment where the power network expects them to "agree, acquiesce and be silent" (Fairclough 1999b:235).

5.0 Conclusion

In this article, we have argued how a power/ideology/language formation enabled acts of deception to be used for personal gain at the cost of democracy. In both Case 1 and Case 2, we attribute the success of deception to the council members' social cognition of the Faculty's discourse which supported the dean's power and teachers' acceptance. The absence of interface—with exception of student representative's reaction in Case 1 and JJ's in Case 2—between council members and the Faculty's discourse is evidence of compliance for tangible reasons: right to study, right to work, salary and promotion—prerogatives which in a democratic environment are taken for granted.

The capacity for language critique (increased critical language awareness, which enables one's participation in discourse and patterns of social power) is usually given to the individual through educational institutions (Hawkins 1984; Fairclough 1999a:220). Suggesting how to construct a "linguistics for the next century" (Candlin 1999:x), CDA contributes, as pointed out by Urry (2000:174), to assuring full cultural participation—in terms of possessing information, representation, knowledge and communication—of all social groups within world society, thus paving their way to "global citizenship." Failing to critique the faculty's discourse, council members neglected their historical task (cf. Fairclough 1999a:220): inculcation of cultural meanings, social relationships and identities, improvement of communicational skills, and, above all, development of capacity for language critique.

Through manufacturing consent, Faculty power holders naturalised the "subject positions" of Faculty staff and students (Fairclough 1999b:105). Not until the Faculty's discourse becomes more transparently related to ideology, will Faculty staff and students be able to transform from "powerless subjects" to "powerful participants" (Fairclough 1999b), and become part of the *power-language* link. Only then, will discourse at the Faculty for Ore Searching move into an associational public space where "differences are brought together" and become "action in concert" (Arendt 1973:56). Emancipation of the Faculty's staff/students will be attained through language, but also manifested in it. This is inevitable, for power goes to those who are "seen to do well" (Kanter 1977), and who is more aware of that than the oppressed?

Notes

[1] Research for this article was realised within a three year (1997-2000) ALIS (Academic Links and Interchange Scheme) project *Hotel & Tourism Management Education Development*. Supported by the British Council and MOST, it was one of a series of research projects aimed at aiding universities in postcommunist transitional countries in developing their courses and curricula.

[2] Withholding information is typical for authoritarian management, who, perceiving knowledge as power & money, will not share it easily. Anybody who insists on free information dissemination is treated as a menace to the organisation's power structure (cf. Davenport 1994; Legge 1995).

[3] In a more participative climate, JJ's suggestion could have been used for creating a programmed conflict, i.e. inviting all Faculty council members to defend/criticise the suggestion on the basis of facts, rather than personalities and individual interests (Kreitner & Kinicki 1992: 376-381).

References

Arendt, H. (1973) *The Human Condition*. University of Chicago Press, Chicago

Candlin, C. 1999. *General Editor's Preface to Norman Fairclough "Critical Discourse Analysis"*. London: Longman, pp. vii-xi

Beaugrande, R. de (1997) *New Foundations for a Science of Text and Discourse: Cognition, Communication, and the Freedom of Access to Knowledge and Society.* Norwood, New Jersey: Ablex Publishing Corporation

Cialdini, R. B. 1984. *Influence: How and why people agree to things.* New York: Morrow

Davenport, T. H. (1994) Saving IT's Soul: Human-Centered Information Management. *Harvard Business Review,* March-April, pp. 41-53

Dickson, D., Saunders, C. & Stringer, M. (1993) *Rewarding people: the skill of responding positively.* London: Routledge

Dika. M. (1998) Foreword—a reviewers' word. In D. Bošnjak, Z. Pi uljan and D. Raj i *An Interpretation of The Law of Higher Education and the Law of Scientific and Research Activity.* Zagreb: Hrvatska sveu ilišna naklada

Fairclough, N. (1995) Media Discourse. London: Edward Arnold

Fairclough, N. 1999a. *Critical Discourse Analysis.* London: Longman

Fairclough, N. 1999b. *Language and Power.* London: Longman

Fairclough, N. (2000) Discourse, social theory, and social research: The discourse of welfare reform. *Journal of Sociolinguistics,* Vol. 4, No. 2, pp. 163-195

Folkes, V. S. 1985. Mindlessness and mindfulness: A partial replication of Langer, Blank & Chanowitz. *Journal of Personality and Social Psychology,* 48, pp. 600-604

Fox, J. (2000) Approaching managerial ethical standards in Croatia's hotel industry. *International Journal of Contemporary Hospitality Management,* 12/1, pp. 70-74

Fox, R. (2001) *Business Communication.* Zagreb: Hrvatska sveu ilišna naklada

Fox, R. & Fox, J. (2001) Transformation and power: the Croatian case. *International Journal of Contemporary Hospitality Management,* 13/1, pp. 43-46

Foucault, M. (1980) *Power/Knowledge.* Brighton: Harvester

Gali , G. (2000) On New Year's Eve every fourth employee on the dole. *Novi list / poslovni,* 20 December 2000

Hawkins, E. (1984). *Awareness of Language: An Introduction.* CUP

Jackall, R. (1996) The Social Structure of Managerial Work. In Barry Castro (ed.) *Business & Society,* OUP

Kanter, R. 1977. *Men and Women of the Corporation.* New York: Basic Books

Kintsch, W. (1989) The representation of knowledge and the use of knowledge in discourse comprehension. In R. Dietrich and C Graumann (eds) *Language Processing in Social Context.* North Holland, Amsterdam, pp. 185-209

Kreitner, R. & Kinicki, A. 1992. *Organizational Behavior,* 2nd Edition. Burr Ridge, Illinois: IRWIN

Lakoff, G. and Johnson, M. (1980) *Metaphors We Live By.* Chicago and London: The University of Chicago Press

Ledi , J., Kova , V. and Rafajac, B. (1998) Students' Position Regarding the Quality of the Teaching Process at the University. *Društvena istra ivanja,* 4-5. pp. 619-637

Legge, K. 1995. *Human Resource Management: Rhetorics and Realities.* London: Macmillan Press Ltd.

Lomnitz, L. A. 1977. *Networks and Marginality.* Academic Press, New York

Marinkovi Škomrlj, E. (2000) Fundamental Academic Principles Destroyed. *Novi List,* 22. May 2000, p. 14

Matkovi , M. (2000) There are not many paupers in Croatia, but 80% of all citizens feel poor. *Vjesnik,* 20. September

2000, pp. 1, 5

Metts, S. (1989) An exploratory investigation of deception in close relationships. *Journal of Social and Personal Relationships*, 6, pp. 159-179

Ng, S. & Bradac, J. (1993) *Power in Language*. Newbury Park: SAGE Publications

Petty, R. E. & Cacioppo, J. T. (1986) *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer Verlag

Redfern, W. (1989) *Clichés and Coinages*. Oxford: Basil Blackwell

Savitt, R. (1995) Critical Issues of Privatization: A Managerial Perspective. In R. Culpan & B. N. Kumar (eds.) *Transformation Management in Postcommunist Countries: Organisational Requirements for a Market Economy*, Westport: Quorum Books

Smith, N. and Vigor, H. (1991) People in Organizations. OUP

Srdo , S. (2000) Our Management is Incompetent and Corrupt. *Novilist*, 22 May 2000, p. 5

Thompson, J. B. (1990) *Ideology and Modern Culture*. Cambridge: Polity Press

Urry, J. 2000. *Sociology beyond Societies*. London: Routledge

van Dijk, T. A. (1996) Social cognition and discourse. In H. Giles and W. P: Robinson *Handbook of language and social psychology*. Chichester: John Wiley and Sons

About the Authors

Renata Fox
Faculty for Tourism and Hospitality Management
University of Zagreb
51410 Opatija, POB 97, Croatia

Email: Renata.Fox@hika.hr
Tel: ++385/51/292-633

Renata Fox (PhD, University of Zagreb, Croatia) is an assistant professor at the Faculty for Tourism and Hospitality Management. Her classroom teaching includes graduate courses in English for Management, German for Management and a postgraduate course in Business Communication. Dr. Fox has lead a project *Hotel & Tourism Management Education Development*. Her research interests lie in the area of communication, critical discourse analysis and cultural studies. She is the author of the book *Business Communication*, and has published in *English for Specific Purposes*, *The International Journal of Contemporary Hospitality Management* and *Revue de tourisme*.

John Fox
Faculty for Tourism and Hospitality Management
51410 Opatija, POB 97, Croatia University of Zagreb

Email: john.fox@inet.hr
Tel: ++385/51/292-633

John Fox (PhD, University of Zagreb, Croatia) is a senior lecturer at the Faculty for Tourism and Hospitality Management, Opatija, Croatia. He has taught English in a number of countries in transition, and has participated on professional courses for the tourist industry. Dr. Fox is co-author of two books: *What an Entrepreneur Should Know* and *Animation in Tourism*. He has conducted research in communication and behaviour in the hospitality industry. His findings have been published in *The International Journal of Contemporary Hospitality Management, Revue de tourisme* and in proceedings of international conferences.

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

66

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

67

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Technical and Ethical Issues in Indicator Systems:
Doing Things Right and Doing Wrong Things

Carol Taylor Fitz-Gibbon
University of Durham

Peter Tymms
University of Durham

Abstract
Most indicator systems are top-down, published, management systems, addressing primarily the
issue of public accountability. In contrast we describe here a university-based suite of "grass-
roots," research-oriented indicator systems that are now subscribed to, voluntarily, by about 1 in
3 secondary schools and over 4,000 primary schools in England. The systems are also being used
by groups in New Zealand, Australia and Hong Kong, and with international schools in 30
countries. These systems would not have grown had they not been cost-effective for schools.
This demanded the technical excellence that makes possible the provision of one hundred percent
accurate data in a very timely fashion. An infrastructure of powerful hardware and ever-
improving software is needed, along with extensive programming to provide carefully chosen
graphical and tabular presentations of data, giving at-a-glance comparative information. Highly
skilled staff, always learning new techniques, have been essential, especially as we move into
computer-based data collection. It has been important to adopt transparent, readily understood
methods of data analysis where we are satisfied that these are accurate, and to model the
processes that produce the data. This can mean, for example, modelling separate regression lines
for 85 different examination syllabuses for one age group, because any aggregation can be shown
to represent unfair comparisons. Ethical issues are surprisingly often lurking in technical
decisions. For example, reporting outcomes from a continuous measure in terms of the percent of
students who surpassed a certain level, produces unethical behavior: a concentration of teaching
on borderline students. Distortion of behavior and data corruption are ever-present concerns in
indicator systems. The systems we describe would have probably failed to thrive had they not
addressed schools' on-going concerns about education. Moreover, data interpretation can only be
completed in the schools, by those who know all the factors involved. Thus the commitment to
working closely and collaboratively with schools in "distributed research" is important, along
with "measuring what matters"... not only achievement. In particular the too-facile interpretation
of correlation as causation that characterized much school effectiveness research had to be
avoided and the need for experimentation promoted and demonstrated. Reasons for the
exceptionally warm welcome from the teaching profession may include both threats (such as the
unvalidated inspection regime run by the Office for Standards in Education) and opportunities
(such as site based management).

If the technical infrastructure is effective, data turn-around quick, data presentation attractive and readily interpreted, then the indicator systems will probably grow and this growth itself demands further technical capabilities, such as running a high-capacity server and creating a central database that can be accessed by researchers and secretaries alike. This central database needs to be relational in order to store efficiently the hundreds of thousands of students with hundreds of variables attached to each student in thousands of schools over many years. It must have an extremely friendly front end, so that secretaries can readily track the mail-out of questionnaires and the return of data, plus a massive invoicing system if individual schools can join the project and pay on their own account. Alternatively, school districts might pay for groups of schools.

Finally the infrastructure needs communication on a regular basis with all schools. Newsletters, a website and conferences are important, particularly as teachers become conference presenters and have a credibility with fellow teachers that researchers lose after some years away from the classroom.

We have been fortunate in working with teachers and headteachers ready to welcome, and make themselves familiar with, streams of data. Some government policies have also helped to make indicator systems important and feasible in the UK: the framework of achievement tests shown in Figure 1, the site-based management legislation requiring school districts to devolve about 80 percent of their budgets to schools, and open enrolment policies allowing parental choice of schools. These were intended both to put schools into competitive situations and also given them some freedom of action derived from having budgetary control.
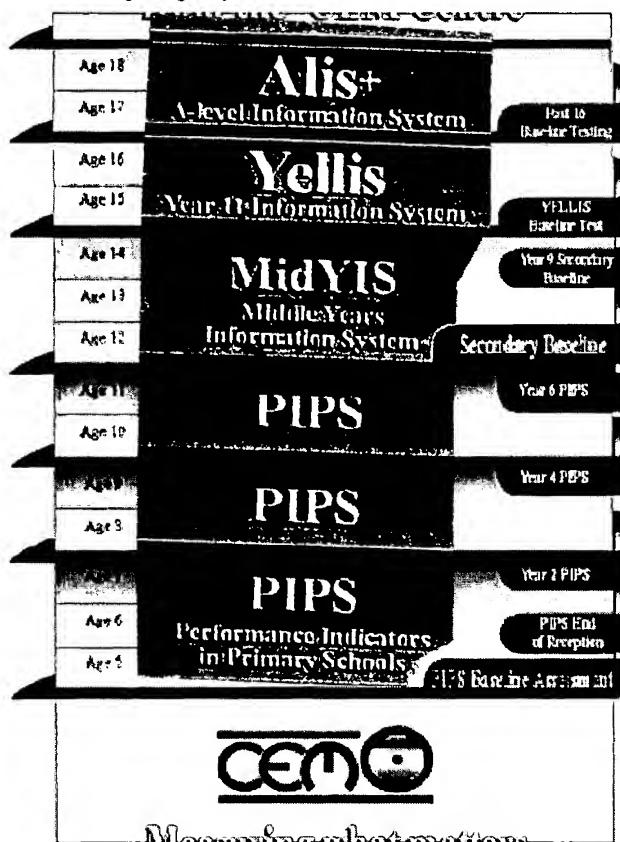
Figure 1. Achievement framework: national tests are provided for students, ages 7, 11, 14, 16, and 18 years.

If the infrastructure for indicator systems can be created, then a cost-effective system is feasible. We now consider the design of such a system, including choosing what to measure, collecting the data, analysing, reporting and interpreting the data.

Choosing indicators

The advice to select a few key indicators is often given (e.g. Lightfoot, 1983 Somekh Convery, Dlaney, Fisher, Gray, Gunn, Henworth and Powell, 1999 p30 and p 34). Whilst this might make life easy, the temptation should be resisted and the advice rejected. A few indicators cannot reflect the complexity of institutions and will undermine the system as gaming takes hold. Given a few indicators, the effort is focused on these concerns alone. Furthermore it is difficult to know which indicators will become important in the future, so that what is now considered to be a key indicator may become of less concern in the future. And who is to decide? Multiple indicators for complex organisations are a fairer representation of the multiple realities within each than is any attempt to assign a single label, whether this label be numerical (e.g. average value added) or verbal (e.g. 'coasting', 'failing').

In addition to prior achievement or baseline measures, are there other important covariates? The best source of information about relevant covariates is not what people write about, but what the data shows. Much is written, for example, of the impact of socio-economic status on achievement, but at the pupil level the correlation is generally about 0.3, thus implying that about 9 per cent of variance in the outcomes will be accounted for by knowing the socio-economic status of the student. In contrast, cognitive measures predict about 50 per cent of subsequent variation. To obtain adequate prediction of subsequent achievement ... and therefore the fairest data for teachers...there is no adequate alternative to a cognitive test.

Affective and social indicators

In addition to the cognitive indicators we need to address the affective and social domains. In Victoria, Australia, there is extensive use of questionnaires to students, to staff in schools and to parents. Currently, in the Curriculum, Evaluation and Management Centre, we concentrate on questionnaires to students, since education is primarily aimed at the students who are in our care for 15,000 hours of compulsory treatment. This concentration on students is also designed to keep the indicator systems lean and efficient and costing as little as possible and obtaining as close as possible to a hundred percent response rates. Students can tell us on questionnaires how much they like school, how much they like an individual subject, whether they feel safe in school, their aspirations for the future, their relationships with teachers, their health, traumas in their lives, how they are taught, and how interesting they find each subject, etc., etc. For children in their first year at school we also ask teachers to rate the children's attention, impulsivity and activity levels.

Does all this amount to too many indicators? Certainly, when schools first join an indicator system, they can feel quite overwhelmed by the amount of data that is returned from a fully developed system. For schools in the first few years of participation 'Keep it simple, stupid' might be a good motto, especially as there is evidence that giving people too much data is de-motivating (Cousins and Leithwood, 1986). However, it would probably be better to give people choice.

We now operate a wide variety of systems of indicators that involve paper- or computer-based tests, as well as 'basic' or 'extended' versions, the latter including hundreds of variables. We are moving towards systems that will involve on-line administration of data collection and permit matrix sampling and the inclusion of choice, by students or other respondents to questionnaires, of the domains in which they would like to express their views and opinions. This will need close attention to the reliability of the data collected. Thus the matrix sampling will use scales as the unit of sampling rather than items.

Having decided on how to measure the outcomes that matter, one is not finished with the creation of indicators. Just as prior achievement predicts subsequent achievement, so prior attitudes will predict subsequent attitudes, and in order to compare like with like we need to use regression analyses and look at the residuals. The prediction appears not to be so strong as in the cognitive area, perhaps due to less reliable measures, but about 25 percent of the variance of final attitudes in secondary schools is usually predictable from knowing intake attitudes.

Process variables

An indicator system consisting of dependent variables with appropriate covariates is a complete indicator system. However, an indicator system is only a step along the way to trying to understand what works, and how schooling can be improved. Consequently, some of our indicator systems include process variables such as descriptions of methods of teaching and learning for which students in the 16-18 age range report the frequency of use.

Process indicators serve to generate hypotheses and most importantly, they stimulate discussion of teaching methods among staff in schools and as such are valuable. The important problems in trying to attribute cause and effect must, however, be continuously emphasised.

Qualitative data: always valued.

As Berliner (1992) argued, qualitative data are powerful. Early in the ALIS project, one school was constantly at the bottom of the set of participating schools on a scale assessing attitude to school. It paid very little attention to this fact but then open-ended questions were introduced into the data collection and students' comments were typed up and made available to the schools. The typing disguised students' handwriting and kept the feedback anonymous. When the school read statements like 'We are treated like fifth formers without uniform', 'Staff are sarcastic', 'I wish I'd gone to another school' this qualitative data had an impact that was immediate and led to a re-design of the provision for subsequent students. Having had that experience, the school then watched the quantitative attitude indicators with more concern and we continue to provide typed-up responses to open-ended questions.

Credible data collection procedures for attitudinal data

We have already described how the cognitive data collection is standardised so that the same procedures are followed in every school. This standardisation of data collection is important in collecting data that can be validly compared from school to school.

70

A particular threat to the validity of attitude data could arise from demand characteristics. If students are being asked if they like the school and whether they get on well with teachers yet teachers are looking over students' shoulders, or if the students feel that their questionnaires will be scrutinised by teachers, then the situation becomes subject to possible pressures and influences that could inhibit honest responding. In the secondary school projects, the tape recording that administers the cognitive test introduces the questionnaire part of the data collection by noting that if there is anything they don't understand they should not raise their hand and ask questions because the teacher cannot come to their desk to help them, since the teacher will be staying at the front of the class in order to avoid seeing the responses on any of the questionnaires. Additionally, students are given plastic envelopes in which to seal their questionnaires.

Of course this procedure requires that students can read the questionnaires and this may not always be the case. If there are non-readers, the questionnaire can be tape recorded and students can be given answer sheets with symbols so that they can listen to the questions on the questionnaire and answer on the answer sheet (Fitz-Gibbon, 1985).

Responding to feedback.

The creation of a monitoring system involves a great many decisions and, as a system grows and there is feedback from the users of the system, there is a need to be responsive and flexible whilst holding firm to fundamental principles. In developing an on-entry assessment for 4-5 year olds the intention was that the data would be kept until the children reached the first statutory assessment three years later. But many reception class teachers suggested that we should assess the students again at the end of their first year at school using an extended version of the on-entry assessment. We now do this on a very wide scale and it has proved to be one of the more important innovations with a number of unseen benefits. (For an analysis of the data see Tymms, Merrell and Henderson, 1997).

Matching individual student records from different sources.

The first task in analysing progress data is to match records from baseline tests to outcome measures. The outcome measures should, of course, be curriculum-embedded, high-stakes, authentic tests that reflect work actually taught and worth teaching in the classroom. The use of a standardised multiple choice measure of reading comprehension, for example, is not likely to be fair to schools since teachers may not be able to influence reading comprehension skills once students can read. In other words, there is a problem of lack of sensitivity to instruction. The matching of data from different sources can only be efficiently done by the use of unique identifiers. These preferably should be identifiers containing check digits and the computing facilities to make sure that no identifier is mis-entered.

Transparent analyses vs. sophisticated statistics such as hierarchical linear models.

Einstein said that everything should be as simple as possible, but no simpler. This is a wise, but very challenging, piece of advice. One cannot know how simple a data analysis can be until one has done both simple and complicated analyses and compared the results with representative sets of real data so that one is looking not only at theoretical models but also at actual magnitudes.

When we won the contract to design a national system of value added indicators, the brief we were given asked for data that was 'statistically valid' and 'readily understood'. These two desiderata could well have been in opposition. We analysed the same data sets using ordinary least squares and multilevel models, and found, as we had found previously, that the average residuals indicating the so-called 'value added' scores for departments or schools, correlated at worst 0.93, and more usually higher, up to 0.99 on the two analyses. Thus it was possible to have the data valid and 'readily understood' by using simple regression. The multi-level analysis, requiring special software and a postgraduate course in statistical analysis, was in contrast to the ordinary least squares analysis that could be taught in primary schools. In our experience in the UK the ordinary least squares analysis can certainly be presented to schools so that most members of staff understand the analysis and can use software to re-analyse data as necessary. This accessibility of the data along with the atmosphere of joint investigation (distributed research) probably helped to encourage acceptance of the indicator systems, unlike the situation that sadly seems to have arisen in Tennessee where a highly ambitious, yearly multi-level analysis was tracking students and teachers (Sanders & Horn, 1995; Baker, Xu & Detch, 1995).

The development of multilevel modelling or hierarchical linear models is admirable, provides efficient calculations and rather different error terms, but to use these procedures in day to day indicator system work is likely to lead to less acceptance of the analysis by teachers. Moreover, it is somewhat akin to applying a correction for relativity when considering the momentum of a moving train: theoretically correct, but in scientific terms, an ill-advised tendency to over-precision.

A recommendation in the Value Added National Project was that prompt initial feedback should be based on very simple value added measures taking account of prior achievement and using ordinary least squares regression methods that any school could adopt and replicate. Then, before any data is made public, statisticians should be given access to the datasets to analyse in numerous sophisticated ways in order to see if any of the analyses makes a difference to particular scores.

Adequate and inadequate statistical modelling.

A method of analysing that *does* make a substantial difference is to consider each subject to have its own regression line, since each subject goes through a particular examining process with a chief examiner and statistical moderation of the marks arrived at by experienced markers working to guidelines. Professor Robin Plackett, winner of two gold medals from the Royal Statistical Society, emphasised in his lectures, usually in his opening sentences that the question to ask, first and foremost, was what processes produced the data. The essence of good statistical modelling is to model the process that produces the data.

From the very start, with the A Level Information System in 1982-83, it was clear that the regression line for mathematics was quite different from the regression line for English and implied that for the same level of prior achievement students would come out with two grades lower taking the Advanced examination in mathematics than they would taking Advanced English (Fitz-Gibbon,1988; Fitz-Gibbon and Vincent, 1994). Regrettably, other researchers (e.g. Donoghue, Thomas, Goldstein, and Knight, 1996) have simply taken the results of all examinations and assumed that the scales could be combined without any adjustment. Having thus confused the data, sophisticated multilevel models were applied to find that there were differential slopes, i.e. slopes that differed for high and low ability intakes. It was even suggested that teachers may be to blame for concentrating on some groups more than others. This was poor data interpretation since a confound (different subjects with different regression lines) was being attributed to teachers' actions without any corroborating evidence.

In Figure 3, we see some of the different regression segments for different subjects based on intake ranges. These indicate very clearly that the intake differs between subjects, that the difficulty level differs between subjects, and that to simply combine the outcome grades as though each subject were of equivalent difficulty is inconsistent with proper statistical modelling based on the processes that produced the data and that the differences are substantial, unlike the difference made by using or not using hierarchical modelling

Segments for A-Levels (1994)

(Segment represents 95% of students)
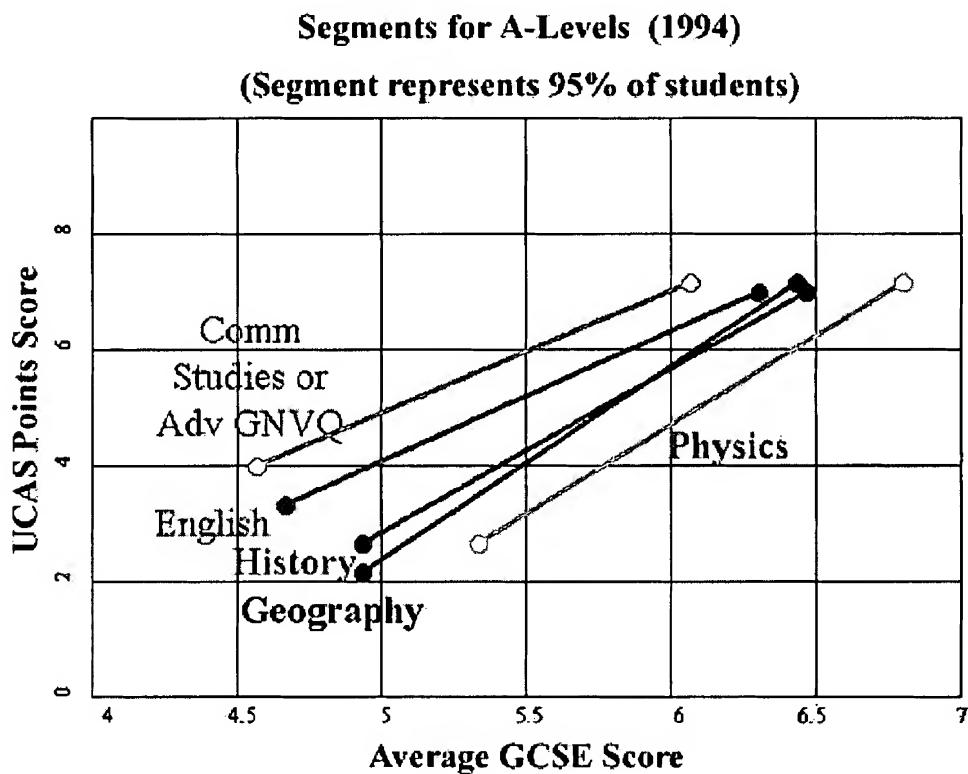
Figure 3. Regression segments showing differences in intake (x-axis) and output (Y-axis) for different subjects

Regression segments, such as were shown in Figure 3 are particularly useful in comparing one subject with another subject, but also in comparing subjects across years. Thus we see in Figure 4 that the average achievement level of the intake is steadily declining (the segment is moving to the left), and the output shows grade inflation (the trend segment is moving up the page). This combination of lower intake range and higher outcome grades has been the pattern with the examinations at age 18 for many years during which time the percentage of students taking these advanced examinations has increased. When these changes are measured against an unchanged baseline, they illustrate the necessary adjustment of 'standards' over time to accommodate expanding range of uptake of advanced courses (Tymms and Fitz-Gibbon, 2001).

ALIS: Chemistry ITDA Regression
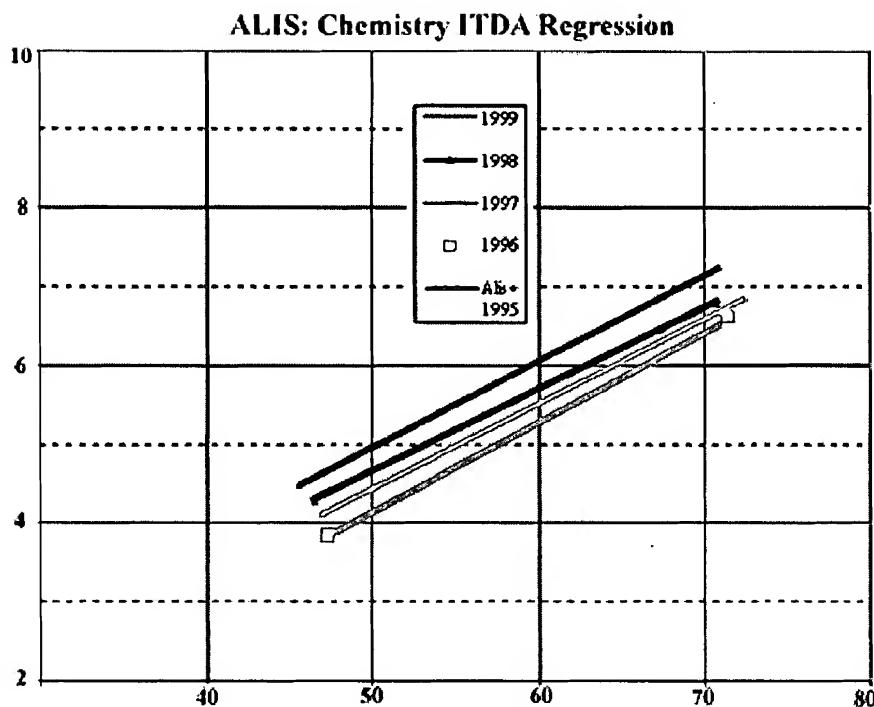
Figure 4. Regression segments for the same subject but different cohorts.

Providing various kinds of feedback, including electronic and web-based feedback.

As with the amount of data, the presentation of data needs to change according to the experience of the school. A school just beginning to get feedback data needs a few clear diagrams and a telephone helpline in case of questions. Schools that have become used to receiving data and have, despite some initial rejection from some departments found it to be useful and credible, start to make more and more use of the data. It therefore becomes valuable to them to have the data provided in Excel spreadsheets, possibly with pre-programmed macros, or in specially prepared software that allows them to undertake procedures such as separating out teaching groups, aggregating by curriculum area, dropping students who have missed substantial amounts of schooling, and adding students for whom data was missing.

Increasingly, as we move from paper-based feedback to sending disks we provide instant feedback. Eventually, with tight encryption techniques this will be directly over the internet.

Chances graphs ... making cognitive tests acceptable.

It has been immensely important in the development of acceptable indicator systems to listen to and to respond to teachers' concerns. It has been important, for example, that baseline tests are not seen as predicting exact outcomes. Fifty per cent of the variation in outcomes is predictable but that means that 50 per cent is not. How can this be represented to teachers who, currently in England, are asked by government agencies to set targets?

This problem was confronted very early in that schools were in some cases preventing students from taking advanced mathematics had they not received a C grade or higher in earlier mathematics courses. When data from a large number of schools was available, in some of which students had been allowed to take the advanced course even having not done well earlier, it was possible to present what we now call 'Chances graphs' (Fitz-Gibbon, 1992, p.288). These graphs show the chances a student had (in retrospect) of getting each grade subsequently. These 'chances' can be represented with simple bar charts showing the empirical percentages of students who actually achieved each grade the previous year. This empirical distribution has great credibility with teachers and students. It is data that actually happened and if it happened once it can happen again. Thus, the low-achieving student is encouraged to recognise that many low achieving students from the previous year well exceeded the average predicted grade for that starting point. By representing their 'chances', we remove the opposition rightly felt to labelling students with single predicted grades and we provide actual data that is motivating for students.

Statistical Process Control Charts (Shewhart, 1986).

A particularly useful representation of the data is one which answers the question 'How is this department doing from year to year, taking into account the number of students in the group and therefore the expected variation in the average from year

73

to year?' Shewhart's brilliant insight into how to represent confidence intervals has proved most useful. By showing the confidence intervals as guidelines to expected variation, data from year to year are very easily scrutinised. Of course, one expects half the results to be above the line and half below the line in some kind of random order. An example of data from a school that might be concerned about its effectiveness is shown in Figure 5 from the A Level Information system.

Departmental Achievement Data

Chart 2.3 Three-year rolling average of standardized residuals
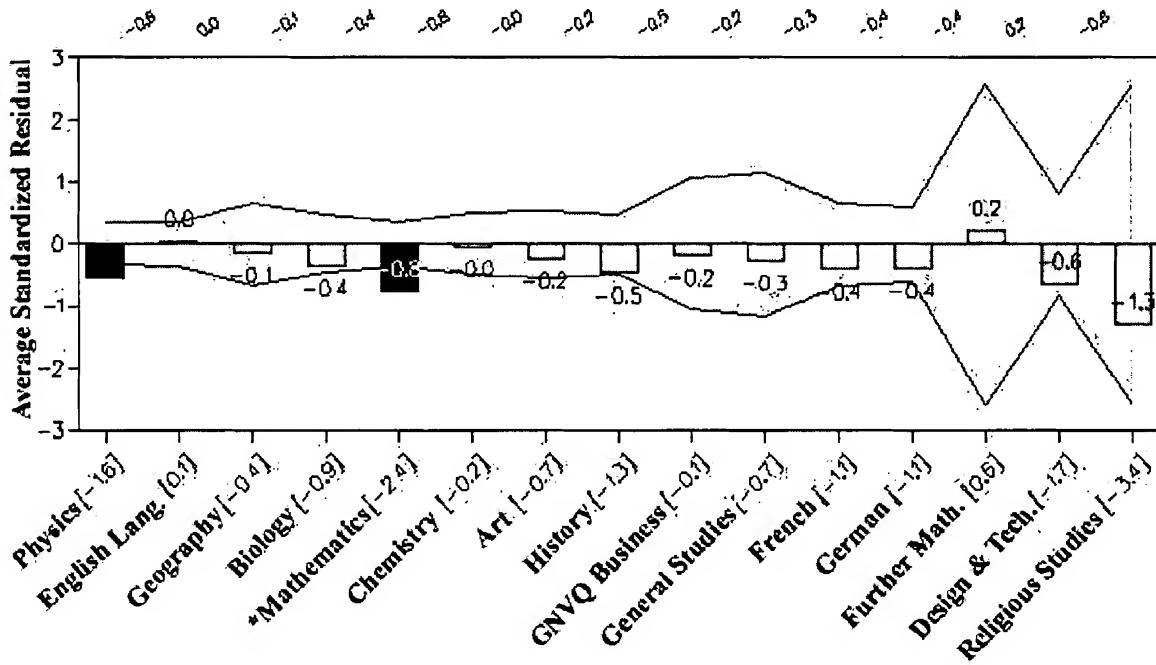
Figure 5. A Statistical Process Control chart for departmental residual gain scores averaged over three years.

The representation involved in statistical process control charts can be applied to presenting the average residuals from various subjects in the same year or using a three year moving average. Each figure is automatically processed from the data in the relational database and a test for statistical significance made. If the variation from zero is statistically significant, the indicator bar turns red so that schools, at a glance, can see which departments are probably doing better or worse than inherent variation. We also warn, however, that statistical significance at any particular level is not a dichotomy between truth and error but simply an indicator on a continuum. The software we provide enables schools to switch easily between a baseline of prior achievement and a curriculum-free baseline.

For publication: the unit of analysis and the unit of reporting.

Compliance with freedom of information legislation and other relevant laws may require that considerable amounts of data are published. The issue as to what should be published is taken up later since it raises ethical issues.

Let it just be acknowledged here that there are issues regarding the reporting unit (we recommend curriculum area, not whole school nor anything finer-grained) but also the problem arises that the vocabulary of research includes words that raise anxieties such as: 'negative', 'below average' and 'regression'. A solution is to show the data in terms of all-round growth with simply variations in the amount of growth. For lay audiences this representation may be more accessible than regression lines.

Interpreting data: Establishing substantive as opposed to statistical significance.

In the statistical process control charts we saw methods of conveying the inherent variability of data samples. It is highly important that politicians and the public recognise that indicators will fluctuate no matter what teachers do. It was commendable that Scotland waited till it had three years' of data before publishing value added measures.

Although we embed statistical significance tests into the data, we also warn schools against using this as a sole criterion. The problems with routine testing at the 0.05 level have been well rehearsed, (Alkin & Fitz-Gibbon, 1975; Carver, 1975; Glass, McGaw and Smith, 1981; Hedges and Olkin, 1985). To assist schools in interpreting the data, we provide both raw residuals

that enable substantive interpretation of differences to be made in the metric in which the examination results are reported[5] and standardised residuals that enable comparisons to be made from year to year. Scales do change; for example in the age 16 examinations, because of grade inflation, an A* was added to the scale as a point above an 'A' grade in the age 16 exams.

Grade inflation due to the standards setting process?

Tymms has suggested that a drift in standards seems to be characteristic in national tests in English primary schools. The reasons for this are connected with the practice of piloting items and setting their difficulty from the results on students who knew they were simply taking an exercise. This 'adrenaline-free' un-prepared testing situation might produce lower performance that would then serve as the benchmark against which the exams were calibrated the following year. Taken under genuine examination conditions, with revision time having been invested and the adrenaline flowing, students might well be producing much better results than those calibrated. Hence the unconscious drift in 'standards'.

Helping users to interpret the data.

It is an unusual teacher-training course that prepares teachers for the kind of information that is now available in England through professional monitoring systems. And yet the information is now seen as vital to many educational professionals. In the course of setting up the CEM Centre projects we have run hundreds of sessions to explain the feedback and to discuss the implications. Further, many courses have been run locally for schools to understand and use the data. Our feeling is that the enormous need for in-service work is an essential part of any monitoring system and that the extent of the need for conferences and workshops often only becomes apparent as the project starts running. It has been standard practice for many years now for our conferences to involve teachers as presenters of the data (e.g. there is a video of a head teacher addressing an early conference and he was a speaker in New Zealand following our involvement there. (Cooper, 1995, video)

Dealing with issues of cause and effect: what works?

This is the most important aspect of data interpretation. It would be wrong to imply to schools that indicator systems are all they need to find out what works. It could take years, even if the search were successful. A school may implement an innovation and the indicator suggests worse results. But perhaps the results would have been even worse without the innovation. Who knows? So the school repeats the innovation and the results stay the same. So another year's data are awaited—and so on.

If instead of this year by year indicator monitoring, if a school joined with 20 other schools and a random 10 implemented one innovation and the other random half implemented a different innovation, all schools would receive 20 years of data in one year. By adopting the methods of science, learning is speeded up and made more reliable.

The fundamental distinction between observation and experimentation must never be blurred.

Epidemiology and clinical trials both have their virtues, but the clinical trials are necessary to establish sound evidence as to what works. That concept applies in education as in medicine and the term 'evidence-based' is now becoming popular. As 'value-added' became the popular word for residuals, evidence-based may become the popular word for experiments. The need not to over-claim for the value of monitoring systems brings U.S. to the next major section of this paper, ethical issues.

Ethical Issues

A major ethical imperative is to do good rather than to do harm. At the very least we might try to observe the Hippocratic oath and 'at least do no harm'. But how do we find out what does harm to students, to society, to academic subjects, to staff?

Evidence of the likely impact of indicator systems on participating schools will be considered including the small number of controlled trials that exist. In addition to this question about the overall impact of indicators there are numerous ethical issues to be addressed that arise in the course of running indicator systems. Each represents a potential source of net harm, a potential negative in a cost benefit analysis.

Some of the questions that arise are:

- Do indicator systems really help schools and affect achievement—or are the admittedly modest funds misspent?
- Should indicator systems lead to a single national, or state, curriculum in order to have a common standard?
- What is the effect of analysing by gender, ethnicity, socio-economic status and religion—does this common activity perpetuate stereotypic thinking?
- What are the effects of poorly chosen indicators, such as those dichotomising continuous data distributions, as in 'percent above x'?
- What are the effects of benchmarking, i.e. comparisons with putatively 'similar' schools?
- Data corruption—does it happen and, if so, who is to blame?
- Is personnel work in public acceptable? (e.g. publishing indicators per teacher)
- Is performance related pay justified?

- Will over-reliance on indicator systems delay the search for better sources of evidence?
- What is the role of the public sector? How can an internal market get the advantages of competition and diversity without the disadvantages of 'the bottom line'? Stakeholders not shareholders?

Do indicator systems really help schools and affect achievement?

It could be argued that because schools freely choose buy into indicator systems this is proof that they find indicator systems useful. However, people buy snake-oil, and the commercial argument is never adequate. People bought treatment with phosphorus that was actually very damaging, and even without a commercial pressure, treatments are provided that do harm simply because adequate evidence has not been collected. What evidence do we have, of a disinterested and objective kind, that indicator systems help schools and, for example, affect achievement?

Cohen (1980) ran a meta-analysis of controlled trials of: no feedback from students to lecturers vs. feedback from students to lecturers vs. feedback from students to lecturers supported by discussions with "an expert." The feedback the same lecturers received in subsequent years improved most in the third condition, and least in the first condition. This result is important. When the ALIS project was about four years old a request was made to a committee at the DfEE (then the Department of Education and Science) inviting them to conduct a randomised controlled trial of the impact of this performance indicator system. The Coopers & Lybrand (1988) report had recommended devolved financing and the use of indictors and the Department was interested. Unfortunately the funds were not found for this potentially important trial. Tymms ran a controlled trial in introducing performance indicators in primary schools into a North Eastern school district in England. A modest effect size (ES_{rct}) of 0.1 was found. This was, however, in 1994 before primary schools were under pressure regarding the publication of examination results and there was no "expert" advice available.

Coe experimented with giving *additional* feedback in the A Level Information System to individual teachers rather than just to school departments. Thus the effect of the randomly assigned feedback was measured not against no feedback but against already substantial feedback, so to expect any further improvement was perhaps optimistic. Nevertheless as a result of giving classroom by classroom analyses to the teachers concerned, rather than simply departmental data from which this information could be extracted, there was an achievement gain of $ES_{rct} = 0.1$ on the high stakes, externally assessed examinations taken at age 18 years[6]. In the Value Added National Project, Tymms experimented with kinds of feedback and found that for primary teachers, tables appeared to be better understood and also, importantly, appeared to have had more impact than graphical feedback. The average Effect Size across English, mathematics and science was $ES_{rct} = 0.2$ (Tymms, 1997, p12).

In the Years Late Secondary Information System[7], a list of under-aspiring students is produced by combining students' intentions regarding continuing in education with their baseline scores. Many schools given the list of under-aspiring students set up mentoring sessions or special monitoring. Unfortunately, good intentions do not guarantee good outcomes (McCord, 1978; McCord, 1981; Dishion, McCord et al, 1999). Aware of our ethical responsibility not to have teachers wasting their time and in order to avoid harming students, we obtained permission from some schools to only feed back to them a random half of the list of their under-aspiring students. In following up these schools and comparing the outcomes of the named under-aspirers versus the unnamed under-aspirers, we actually have found more differences in favour of the unnamed group than the named group. Indeed, naming students resulted in an overall effect on examination progress, adjusted for prior achievement of value added decrement of $ES_{rct} = -0.38$. Naming seemed to have little effeet on whether or not students were counselled at all ($r = 0.01$) but the more counselling sessions that any students, named or not, received the worse were their value added scores ($r = -0.22$). Only 15 schools were involved in this first experiment, but it calls into question many facile beliefs about how achievement can be improved. The findings are challenging and the experiment is being repeated with thirty schools. It illustrates how an indicator system can move the profession forward to proper experimentation.

Should indicator systems lead to a National Curriculum?

The resistance to a National Curriculum in the U.S. has contributed to the slow development of curriculum-embedded, high-stakes, authentic tests. In England, where external curriculum-embedded assessments have been used for decades and school performance tables are published using raw results, moves have been made towards value added systems. These will increase the high stakes nature of the external examinations and, at the same time, government pressure on the Qualifications and Curriculum Authority has led to a reduction from seven independent examination boards to three conglomerates of the former boards. Furthermore there has been a reduction in the number of syllabuses on offer for secondary schools.

Meanwhile in primary schools, a single National Curriculum has been imposed and all primary students sit the same tests designed to the same syllabuses at the ages of 7, 11 and 14 years. The specification of a National Curriculum concentrating on particular subjects and the publication of these data has put schools under pressure to drop attention to such areas as the fine arts, the performing arts, and physical education, and to concentrate on those indicators that are published. All schools are forced to do the same curriculum unless exemptions are granted.

This restriction and concentration certainly represents a downgrading of the professional status of teachers who can now make few important decisions, and it may contribute to declining levels of satisfaction of teachers. At the very least there should be various curricula available to be chosen, as was the case for decades for teachers of students aged 16 and 18 years.

Thus, a teacher who preferred to teach physical geography rather than economic geography could find a syllabus in which the proportion was attractive for that teacher. Another reason for maintaining choice and diversity in syllabuses is that in the entire population a much broader range of skills is thereby likely to be developed. Choice and diversity also keep the examination boards in competition and this ought to lead to an improvement in the quality of the service that they provide. Unfortunately, since they have a virtual monopoly endowed by government approval, it will not be likely that examination boards drop their poor practice unless required to do so. Examples of poor practice from examination boards are leaving the names of students and their schools on the examination paper when it is being assessed. The name of the pupil and the school will often contain clear evidence regarding the pupil's gender, ethnicity, social class and religion. In the face of this information, can essays be read in a totally unbiased way? Further poor practice is the lack of provision of inter-marker reliability data (Fitz-Gibbon, 1996, p.115).

What is the effect of analysing by gender, ethnicity, socio-economic status and religion?

There may be differences between groups, but ethnicity is very poorly defined; socio-economic status is not well-measured; and neither of these variables, is alterable by the school. Alterable variables (Bloom, 1979, 1984) are the key to improvement and accountability. Religion is perhaps an alterable variable, but if we find Catholic schools are doing better than Protestant schools, do we draw the inference that we should make schools turn Catholic? Or vice versa? The habit of analysing by these unalterable variables may simply be a result of the pressure to produce academic papers, whether they contribute to practical or theoretical developments or not. Given a body of data it is easy to break it down by these categories, and report the differences. The fact that it leads nowhere has not been a major consideration in social science research.

The fact that such analyses perpetuate stereotyping should also be a matter of ethical concern. That these correlational analyses do not promote the search for strong evidence as to what works, is certainly a matter for ethical concern. Attention should be directed towards alterable variables rather than unalterable categories into which human beings are grouped, which is the first step to stereotyping. These analyses become particularly a matter of concern when teachers are presumed to be somehow to blame for the 'under-achievement' of boys at the age of 16 as compared with the achievement of girls. Group differences make catchy headlines in the newspapers. While there may sometimes be a need to track group differences, there is a more important need to educate users of data about the size of the effects being studied and what is known about altering the situation. Boys are smaller than girls at age 11. Should they be stretched? Are teachers responsible?

The use of a "percentage greater than" criterion in reporting

The most egregious mistake made in performance data in England has been the DfEE's [8] introduction of arbitrary dichotomies into continuous data. Thus, primary school students' achievements are publicly reported in terms of the percent of students in each school above a certain level, called Level 4. This has the unfortunate implication that students below Level 4 have in some way failed their school or failed in their schooling. This is extremely unethical, since for some students a Level 4 achievement is an excellent achievement, whereas for others a Level 4 is a failure to reach their potential.

Furthermore, to draw an arbitrary line through a continuous outcome data almost always leads to very negative reactivity. At the secondary level the damaging and unethical impact is a concentration on D students because the reporting line is the percentage of students getting Grade C or above. Time, effort and money have been spent on D students to the neglect of more able and less able students.

If, on the other hand, an average points score is used as the outcome measure, the implication is to work with each pupil to obtain their maximum performance. This is ethical behaviour, it is the kind of behaviour teachers wish to adopt, but it is made impossible by the reporting of indicators based on arbitrary dichotomies in the data.

The effects of arbitrary benchmarks

In England, official bodies such as the Office for Standards in Education, lacking pupil level value added measures, compare schools with 'similar' schools. The classification of 'similar' is usually made on the basis of the percent of students receiving free school meals. However, two schools can both have 20 per cent of students receiving free school meals but otherwise have quite different profiles. For example, one may have a larger proportion of children who also come from schools with very high levels of achievement. Such a school benchmarked against a school with the same percent of free school meals will look very good at the expense of the other school, but the comparison is spurious. Such benchmarking is an inadequate way of making comparisons. The only fair comparisons are with similar students in other schools. There are no similar schools.

It is certainly not ethical to make unfair comparisons which in some cases carry financial consequences for the institution concerned and can lead to job losses and demoralisation. Indeed, to take a most extreme and serious consequence, Ofsted inspectors rely on poor benchmarking data and also sit in classrooms judging teachers. Ofsted inspections have recently been cited in four inquests following suicides by teachers (Times Educational Supplement, April, 2000).

Fair data carefully interpreted is a defence against the inequities of the Ofsted system, problems reported at length to a

Select Committee of the House of Commons (website: http://www.cem.dur.ac.uk/) (Kogan, 1999; Fitz-Gibbon, 1998; Fitz-Gibbon and Stephenson, 1999).

Data corruption: when does it happen and who is to blame?

In an article entitled 'On the unintended consequences of publishing performance data in the public sector' Peter Smith, Professor of Economics at the University of York, identified a *'huge number of instances of unintended behavioural consequences of the publication of performance data'* (Smith, 1995). He named eight problems associated with non-effective or counter-productive systems:

- tunnel vision;
- sub-optimisation
- myopia;
- measure fixation;
- gaming;
- ossification;
- misinterpretation;
- misrepresentation.

These can be seen as distortions of behaviour and attention (the first six) and data corruption (the last two). With the sole exception of ossification, every one of these possibilities was raised by headteachers in open-ended items in the questionnaires used in the Value Added National Project. Thus in education these are not theoretical problems but actual, already-perceived problems (Fitz-Gibbon, 1997).

W Edwards Deming (1986) warned that "When there is fear we get the wrong figures." In primary schools in England there have been instances of teachers opening the examination papers the week before assessments and making sure that students were well-prepared. This unfortunately has negative consequences for the school subsequently, since higher than reasonable achievement levels will be expected.

A more subtle form of data corruption is to exclude students who are not going to produce good examination results. In England following the advent of publication of raw achievement levels in the form of 'School Performance Tables',[9] exclusion rates increased 600 per cent. Exclusions from school may be the beginning of an increased risk of delinquency, drug-taking and criminality—is this a price worth paying for the publication of school performance data? It is widely acknowledged that there was a causal link here: schools saw a way to improve their standing in the tables and excluded difficult students. The government some years later responded by publishing exclusion rates and making an issue of 'inclusion'... but the impact had already taken place for many students.

As further pressures arise from 'performance management' (performance related pay systems) it may not be long before we see baseline measures declining so that value added measures look better, particularly when old IQ tests are used for baselines and are not standardised in their administration procedures.

Personnel work in public

Whole school indicators should be avoided because the evidence is that there is more variation within a single school than is generally found between schools. Furthermore, the use of whole school indicators encourages the rank ordering of schools and the public is not prepared to interpret rank orders adequately. Very small differences in the indicator can move a school through many positions in a rank ordering in the middle of a distribution. To avoid simple rank ordering, schools were sometimes put into bands, but this too can be damaging if bands A through E are used. Schools in 'D' and 'E' bands are castigated but in any distribution half have to be below average. This may be politically unpalatable but such is the nature of the average.

If indicators were published for each teacher, this would be tantamount to doing personnel work in public and would be unacceptable. And yet data cannot be withheld from the public unreasonably, so some compromise is needed: not whole school indicators and not individual teacher indicators.

The compromise recommended in the Value Added National Project was to use *curriculum area* as the unit of reporting. This has the virtue of enabling parents to look for schools that seem to be doing well in the area in which their children are most interested (e.g. performing arts or mathematics and science curriculum areas). Of course, in small schools there may be no distinction between the indicators for a curriculum area and for a teacher. There needs to be some restriction put on the size of sample that can be reported publicly. The CEM Centre is developing these indicators for the provision of data at the LEA[10]/School District level as opposed to individual school level, where the data is presented department by department for affective and cognitive indicators, and student by student in the cognitive area. Within the individual school, further analyses can be undertaken to obtain data teacher by teacher. Such analyses are made easy by our provision of the school's data in software packages called Pupil Assessment and Recording Information System (PARIS).

Performance related pay

George Soros, in his book *The Crisis of Global Capitalism*, elaborates on his concept of reflexivity. His point is that, in the social world, where perceptions can influence behaviour, saying 'it is so' may indeed 'make it so'. Mistaken beliefs about the nature of the physical world have no influence on the physical world, but distortions of beliefs about the social world can have an impact. One of the distortions promulgated by those seeking to implement performance related pay is that pay is the great motivator. This is only a hypothesis, and before huge amounts of money go into implementing performance related pay systems, they should be put to an experimental test in which some schools get performance related pay and other schools get equivalent money to spend as they wish.

The negative influences of performance related pay are potentially the destruction of team work, the demoralisation of those who do not get a performance pay rise, the corruption of data due to the chance to make financial gain from 'good' exam results, and the message sent to students that teachers work for pay: not for their love of the subject, not for their concern for their students, but for pay. According to Soros's concept of reflexivity, this very implication can make itself come true as beliefs can be distorted.

Will over-reliance on indicator systems delay the search for better sources of evidence?

Just as epidemiology is inadequate as a basis for assessing medical treatments, so indicators are inadequate as a means of establishing 'what works' in education. As argued earlier, as schools experience the yearly receipt of indicators of the progress of every student and see the data accumulating in Statistical Process Control charts, they realise that simply watching the indicators, whilst very important, is a slow way to find out 'what works'.

The launch, in Philadelphia in February 2000, of the 'Campbell Collaboration' represents a major effort to create a more just and effective society. It is important that the provision of indicators will support this important step forward and they do, indeed, provide an excellent context in which to conduct experiments: by embedding experiments in institutions with on-going indicator systems, time series data with randomised interventions becomes a very powerful source of high quality evidence.

The role of the public sector

Indicator systems, feasible because of computers, may make the public sector, and in particular public sector management, a fascinating exercise in applied social science. Finally social scientists may have some responsibility for more than arguments and papers. The actions of managers and administrators should be guided by social science findings. They can study their success in applying the findings by watching the indicators as business managers watch the bottom line or the share price. Perhaps indeed the pensions of Chief Education Officers could be tied to the long-term outcomes of the students who are in their care for about 15,000 hours of compulsory treatment. However, the public sector, including universities, will need to permit innovation, flexibility, and devolved 'site-based management' and public servants will need to reduce drastically time-serving hierarchies and inefficient bureaucracies.

Conclusion

The most important aspect of an indicator system is its reactivity: the impact it has on behaviour in the system being monitored. All the issues raised above need attention to create indicator systems in which the benefits outweigh the costs.

Porter (1988) described the tensions in how indicator systems may be used. When a headteacher[11] said that our indicator systems had 'Introduced a research ethos into the school' we felt this was exactly what was desirable and ethical. But there are pressures to make indicators part of an aggressive management culture, including target setting and performance related pay. Without knowledge of cause, effect and magnitudes of effects this is likely to be unproductive gaming. Good management requires good science, including the recognition of our ignorance concerning many aspects of schooling. An 'Evidence-Based Education Network' is one of the ways in which we wish to promote the research agenda in our 'distributed research' with schools. The questions are not 'Who is to blame and who needs to be rewarded?' but 'What do we know and how do we find out what works?' A research ethos.

Notes

[1] Each summer, with a turn-around time of a few weeks, the CEM Centre processes hundreds of variables and matched pre-post scores on over a million students. Staff look after 12 servers and a relational database management system (RDMS) used by researchers, secretarial and administrative staff.

[2] The examination system in England has long delivered authentic, high stakes, curriculum-embedded tests, called 'examinations'. The complex authentic tests are based on syllabuses to which teachers teach. The examination papers are published each year along with comments from examiners. The systems were set up by universities. Teachers are hired to mark the authentic scripts to clearly designed criteria. The examinations are 'high stakes' but not punitive but aiming to provide certification that assists in gaining university entrance and jobs.

[3] Further discussion of the statistical issues is available in the Vernon Wall lecture on the website

www.cem.dur.ac.uk/software/.

[4]Roughly comparable to Advanced Placement in the U.S. Advanced level examinations in England are taken at age 18 and there is, for 2001, a new examination the year before.

[5]E.g. 'Levels' in primary schools and 'grades'... A, B C etc. ... in secondary schools.

[6] (To assist readers in distinguishing correlational from experimental findings the ES is subscripted 'rct' if it arises directly from the manipulation in a randomised controlled trial. This practice, (recommended in Fitz-Gibbon, 1999, p. 37) could make meta analyses considerably easier to conduct, especially for electronically published articles.)

[7]YELSIS also known as YELLIS, Year 11 Information System.

[8]Department for Education and Employment, based in London.

[9]WEBSITE: **Error! Reference source not found.**

[10]Local Education Authority, i.e., School District.

[11]Keith Nancekievil, Gosforth High School, Newcastle upon Tyne, England

References

Alkin, M.C. & Fitz-Gibbon, C. T. (1975) Methods and Theories of Evaluating Programs., *Journal of Research and Development in Education, 8*(3), pp. 2-15.

Baker, A.P., Xu, D. and Detch, E. (1995) The measure of education: a review of the Tennessee value added assessment system. Nashville, TN, Office of Education Accountability, Tennessee Department of Education, Comptroller of the Treasury.

Berliner, D.C. (1992) Telling the stories of educational psychology, *Educational Psychologist, 27*(2), pp. 143-161.

Bloom, B.S. (1979). Alterable variables: the new direction in educational research. Edinburgh: Scottish Council for Research.

Bloom, B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as one-to-one Tutoring. *Educational Researcher,* June/July: 4-16.

Carver, R.P. (1975). The Coleman Report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12*(1): 77-86.

Cohen, P.A. (1980). Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-analysis of Findings. *Research in Higher Education, 13*(4): 321-341.

Coopers and Lybrand, (1988). Local Management of Schools. London: HMSO.

Cousins, J.B. and Leithwood, K.A. (1986). Current Empirical Research on Evaluation Utilization. *Review of Educational Research, 56*(3): 331-364.

Deming, W.E. (1986). *Out of the Crisis: Quality Productivity and Competitive Position.* Cambridge University Press.

Dishion, T.J., McCord, J. et al. (1999). When Interventions Harm. *American Psychologist, 54*(9): 755-764.

Donoghue, M., Thomas, S., Goldstein, H. and Knight, T. (1996). DfEE Study of Value Added for 16-18 Year Olds in England. London: DfEE.

Fitz-Gibbon, C.T. & Stephenson-Forster, N.J. (1999). Is Ofsted helpful? An evaluation using social science criteria, in C. Cullingford (Ed.), *An Inspector Calls: Ofsted and its effect on school standards* (London, Kogan Page).

Fitz-Gibbon, C.T. (1985). Using audio tapes in questionnaire administration. *Research Intelligence, 19*:8-9.

Fitz-Gibbon, C.T. (1988). Recalculating the standard. 26-8-88, *The Times Educational Supplement*, p 15

Fitz-Gibbon, C.T. (1990). Performance Indicators: a BERA Dialogue. Clevedon, Avon: Multi-lingual Matters.

Fitz-Gibbon, C. T. (1992). The Design of Indicator Systems - The Role of Education in Universities, and the Role of Inspectors/Advisers: A Discussion and a Case Study. *Research Papers in Education—Policy and Practice, 7*(3), 271-300.

Fitz-Gibbon, C.T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness.* London: Cassell/Continuum.

Fitz-Gibbon, C.T. (1997). *The Value Added National Project: Final Report: Feasibility studies for a national system of Value Added indicators.* London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C. T. (1999). Education: High Potential Not Yet Realized. *Public Money & Management: Integrating Theory and Practice in Public Management 19*(1): 33-40.

Fitz-Gibbon, C.T. and Kochan, S. (2000). School Effectiveness and Education Indicators. In C. Teddlie and D. Reynolds (Eds) *The International Handbook of School Effectiveness Research.* London: Falmer Press, pages 257-282.

Fitz-Gibbon, C.T. and Stephenson-Forster, N.J. (1999). Is Ofsted helpful? An evaluation using social science criteria. In C. Cullingford (Ed) *An Inspector Calls: Ofsted and its effect on school standards.* London: Kogan Page, pages 97-118.

Fitz-Gibbon, C.T. and Vincent, L. (1994*) Candidates' Performance in Public Examinations in Mathematics and Science.* London: School Curriculum and Assessment Authority (SCAA).

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, CA:, SAGE Publications).

Green, D.R. (1974) *The Aptitude Achievement Distinction.* Monterey, CA: McGraw-Hill.

Hart, P.M., Conn, M. & Carter, N. (1992) *School Organisational Health Questionnaire: Manual.* Melbourne, Victoria, Australia, Department of School Education).

Hedges, L.V. & Olkin, I. (1985) *Statistical methods for meta-analysis.* New York: Academic Press.

Kogan, M. (1999) The Ofsted System of School Inspection: An independent evaluation. A report of a study by The Centre for the Evaluation of Public Policy and Practice and Helix Consulting Group. CEPPP, Brunel University.

Lightfoot, S. L. (1983). *The Good High School: Portraits of Character and Culture.* New York, Basic Books.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33,* 284-289.

McCord, J. (1981). Considerations of Some Effects of a Counseling Program. In S.E. Martin, L.B. Sechrest and R. Redner (Eds.), *New Directions in the Rehabilitation of Criminal Offenders.* Washington, DC: National Academy Press, 393-405.

OECD (1998). *Education at a Glance OECD Indicators.*

Porter, A. (1988). Indicators: Objective Data or Political Tool? *Phi Delta Kappan,* (March 1988): 503-508.

Sanders, W.L. and Horn, S. (1995). An Overview of the Tennessee Value-Added Assessment System (TVAAS)—Answers to Frequently Asked Questions. Knoxville, Tennessee: The University of Tennessee.

Shewhart, W.A. (1986). Statistical Method from the Viewpoint of Quality Control. (Graduate School, Department of Agriculture, Washington, 1939). Dover.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration.* 18(2/3) 277-310

Somekh, B., Convery, A. et al. (1999). Improving College Effectiveness: raising quality and achievement. London: Further Education Development Agency.

Tymms, P.B. (1997). The Value Added National Project – Technical Report: Primary 3, London: SCAA.

Tymms, P.B. (1999). *Baselines Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-Added Indicators.* London: David Fulton.

Tymms, P.B. and Fitz-Gibbon, C.T. (2001). Standards, Achievement and Educational Performance, 1976-2001: A Cause for Celebration? In R. Phillips and J. Furlong (Eds.), *Education, Reform and the State: Politics, Policy and Practice, 1976-2001.* London: Routledge.

Tymms, P.B., Merrell, C. and Henderson, B. (1997). The First Year at School: A quantitative investigation of the attainment and progress of pupils. *Educational Evaluation and Research, 3*(2): 101-118.

About the Authors

Professor Carol Taylor Fitz-Gibbon
Director, The Curriculum, Evaluation and Management Centre
University of Durham
England, UK

Email: Carol.Fitz-Gibbon@cem.dur.ac.uk

After 10 years of teaching physics and mathematics in a variety of schools in the U.K. and then the U.S., Carol Fitz-Gibbon conducted a study for the U.S. Office of Education on the identification of mentally gifted, inner-city students and then became a Research Associate for Marvin C. Alkin, at the Center for the Study of Evaluation, UCLA. She completed a Ph.D. in Research Methods and Evaluation, obtained a grant on the design of compensatory education, co-authored a series of textbooks and returned to the UK in 1978 planning to continue work on Cross-age and Peer Tutoring. But the success of an indicator system she developed with 12 schools in 1983 led to other areas. Much of this work is described in the prize-winning book *Monitoring Education: Indicators, Quality and Effectiveness* (1996).

Professor Peter Tymms
Director of the PIPS Project
The Curriculum, Evaluation and Management Centre
University of Durham
England, UK

After taking a degree in natural sciences, Peter Tymms taught in a wide variety of schools from Central Africa to the north-east of England before starting an academic career. He was "Lecturer in Performance Indicators" at Moray House, Edinburgh, before moving to Newcastle University and then to Durham University, where he is presently Professor of Education. His main research interests are in monitoring, assessment, school effectiveness and research methodology. He is Director of the PIPS project within the CEM Centre, which involves monitoring the progress and attitudes of pupils in about 4000 primary schools. He has published many academic articles, and his book *Baseline Assessment and Monitoring in Primary Schools* has recently appeared.

BEST COPY AVAILABLE

other vols. | abstracts | editors | board | submit | comment | subscribe | search

c

83

Education Policy Analysis Archives

¿Exito en California?
A Validity Critique of Language Program Evaluations and Analysis of English Learner Test Scores

Marilyn S. Thompson
Kristen E. DiCerbo
Kate Mahoney
Jeff MacSwan
Arizona State University

Abstract
Several states have recently faced ballot initiatives that propose to functionally eliminate bilingual education in favor of English-only approaches. Proponents of these initiatives have argued an overall rise in standardized achievement scores of California's limited English proficient (LEP) students is largely due to the implementation of English immersion programs mandated by Proposition 227 in 1998, hence, they claim *Exito en California* (Success in California). However, many such arguments presented in the media were based on flawed summaries of these data. We first discuss the background, media coverage, and previous research associated with California's Proposition 227. We then present a series of validity concerns regarding use of Stanford-9 achievement data to address policy for educating LEP students; these concerns include the language of the test, alternative explanations, sample selection, and data analysis decisions. Finally, we present a comprehensive summary of scaled-score achievement means and trajectories for California's LEP and non-LEP students for 1998-2000. Our analyses indicate that although scores have risen overall, the achievement gap between LEP and EP students does not appear to be narrowing.

Education policy concerning the instruction of limited English proficient (LEP) students in the United States has been debated for a number of decades and considerable attention has been given to the best method of instruction for these students. In recent years, the controversy regarding how to best educate LEP students has surfaced in the form of political legislation. According to the United States Department of Education (1994) the term "limited English proficient" refers to individuals who (1) were not born in the U. S. and whose native language is other than English, or (2) come from environments in which a language other than English is dominant. The education of LEP students is important, especially given that during the 1996-1997 academic year U. S. school districts reported an enrollment of approximately 3.5 million limited-English proficient (LEP) students, accounting for 7.4% of the total reported enrollment (Macías, Nishikawa, & Venegas, 1998). The proportional rate of increase in LEP students from 1995 to 2020 is projected to be 96%, compared to an expected increase of 22% for native-English speakers (Campbell, 1994).

The controversy regarding the education of LEP students generally focuses on the amount of instruction provided in

84

the children's native language. English immersion programs provide instruction almost exclusively in English, which the teacher attempts to make accessible to LEP students. Bilingual education programs provide a substantial amount of content area instruction in the students' native language, while some time each day is spent developing English skills. It should be noted, however, that the actual implementation of programs varies across states, districts, schools, and even classrooms (August & Hakuta, 1998; Berliner, 1988).

Proponents of bilingual education argue that without support in their native language, LEP students will fall behind academically while they are learning English (Crawford, 1999; Krashen, 1996). They also argue that if students first learn to read in the language in which they are fluent, they can then transfer those skills to reading in English (Krashen, 1996). Proponents of English immersion argue that instructional time devoted to intensive learning of English will more likely benefit children's academic achievement in a second language environment (Rossell & Baker, 1996).

Recently, the debate surrounding the education of LEP students has shifted to the political arena. Several states, including California, Colorado, and Arizona, have faced ballot initiatives that propose to restrict the types of educational methods and programs that may be used to instruct LEP students. Specifically, these restrictions functionally eliminate bilingual education programs in favor of an English immersion approach. The first such proposition was California's Proposition 227, which passed by a majority vote in 1998. In November of 2000, voters in Arizona approved a similar, but even more restrictive measure, Proposition 203. In early 2001, measures similar to these propositions were introduced in the state legislatures of Massachusetts, Oregon, and Rhode Island.

In this article, we first provide an overview of the media coverage surrounding the implementation and evaluation of California's Proposition 227 and then review scholarly analyses related to its initiation, including both qualitative studies of implementation and quantitative evaluations of student achievement scores. We then discuss several methodological problems we have observed with the use of California Department of Education's Standardized Testing and Reporting (STAR; California Standardized Testing and Reporting, 2000) data to support arguments about the effects of Proposition 227. We limit our discussion to Stanford-9 scores released in 1998, 1999 and 2000 because we are concerned with the validity of claims about the success of Proposition 227 which purport to derive from these specific data. See Hakuta (2001) for remarks on the 2001 Stanford-9 scores of California's English learners. We frame these problems in the context of specific threats to validity and inappropriate approaches to data analysis. Finally, we present a comprehensive reanalysis of the STAR data and summarize achievement trajectories for LEP and non-LEP students. In interpreting our reanalysis, we discuss policy issues we feel cannot be adequately examined based on these data.

All the News That's Fit to Print?
Media Accounts of California's Stanford-9 Scores and Proposition 227

California implemented Proposition 227 during the 1998-1999 school year. During the previous year, California also began statewide administration of the Stanford Achievement Test, 9th edition (Stanford-9). These test results are publicly available, aggregated by grade level for each school, through the STAR system (California Standardized Testing and Reporting, 2000). In the past two years, many educators, media sources, and political stakeholders have reported summaries of these data as evidence of the effectiveness of Proposition 227.

The *New York Times*, which is widely recognized as one of the most influential newspapers in the United States, published a news story focusing on Stanford-9 achievement scores of LEP children in California on August 20, 2000. *Times* reporter Jacque Steinberg claimed that the increase in scores "at the very least" represented "a tentative affirmation" of the vision of Ron Unz (Steinberg, 2000, A1), who had sponsored the California initiative that banned bilingual education two years earlier. The *Times* story appeared as front-page news, running 1,744 words in length, and opened with the following statement:

> Two years after Californians voted to end bilingual education and force a million Spanish-speaking students to immerse themselves in English as if it were a cold bath, those students are improving in reading and other subjects at often striking rates, according to standardized test scores released this week. (p. A1)

Steinberg concluded the test results provide tentative evidence that Proposition 227's prescribed "cold bath" of English immersion is responsible for the increase in scores, and characterized the results as "remarkable." The *Times* piece also included an extensive anecdote of a school superintendent from the Oceanside district who converted from an advocate of bilingual education to a proponent of structured English immersion.

To present a contrasting view, Steinberg included a 59-word paragraph in which he suggested alternative explanations for the increase in scores, citing class-size reduction in particular. However, these alternatives were introduced by the suggestion that Proposition 227 was at least in part responsible for the increase, which Steinberg found to be "remarkable given predictions that scores of Spanish-speaking children would plummet" (p. A1). Steinberg also quoted Stanford Professor Kenji Hakuta, who had conducted an analysis of the test scores and posted them on the World Wide Web the same day they were released. Rather than discussing Hakuta's study, Steinberg briefly summarized that it was Hakuta's view that "few conclusions could be drawn from the results, other than that 'the

85

numbers didn't turn negative,' as many had feared" (p. A1). Steinberg appeared to use Hakuta's quote essentially to sustain his main point, namely, that the increase in test scores is a tentative affirmation of Proposition 227, and a clear indication that educators were wrong to predict children would suffer.

The *Times* story was syndicated in the *Milwaukee Journal Sentinel* (Steinberg, 2000b, p. 3A), where the opposing viewpoint was cut by half, and in the *Baltimore Sun* (New York Times New Service, 2000, p. 3A) and Cleveland's *Plain Dealer* (Steinberg, 2000c, p. 21A), where it was entirely eliminated. After the *New York Times* story appeared, the idea that California's Stanford-9 gains for LEP students resulted from the implementation of Proposition 227 was cited in 24 major U. S. newspapers, frequently without any question of the accuracy of the claim. Of these, 17 (or 71%) gave no voice to opposing viewpoint at all (see Table 1). (Note 1)

Table 1
News Stories in Major Newspapers (Aug. 20, '00—June 9, '01)
Mentioning the Increase in California's Stanford-9 Test Scores as
Evidence of the Success of Structured English Immersion (Proposition 227)

Newspaper	Date of publication	Length of article (in words)	Length of opposing view (in words)
The Plain Dealer	8/20/00	712	0
Milwaukee Journal Sentinel	8/20/00	839	41
The Baltimore Sun	8/20/00	848	0
New York Times	8/20/00	1744	78
The Arizona Republic	8/22/00	679	46
The Christian Science Monitor	8/23/00	1172	0
The Houston Chronicle	8/28/00	261	105
Star Tribune	8/28/00	540	0
USA Today	8/28/00	996	0
Newsday	9/09/00	450	0
The Arizona Republic	9/22/00	844	0
The Christian Science Monitor	9/27/00	862	0
The San Diego Union Tribune	10/06/00	574	0
The Arizona Republic	10/29/00	1283	0
Los Angeles Times	11/07/00	98	46
The Arizona Republic	11/08/00	678	0
New York Times	11/15/00	600	0
The Boston Globe	12/31/00	1125	0
The Atlanta Journ. & Constitution	1/04/01	647	0
The Boston Globe	1/14/01	436	0
The Arizona Republic	3/01/01	431	0
The Arizona Republic	3/02/01	448	40
The Denver Post	3/28/01	811	29
New York Times	4/01/01	228	0
Averages:		721.1	16.0

Interestingly, the Associated Press (AP), which writes stories circulated to its 1,550 clients, including the *New York Times*, wrote a considerably more balanced story a week before the publication of the *Times* piece. AP reporter Jennifer

Kerr's story opened as follows: "Two years after voters ended most bilingual education in California, statewide test scores for non-English speakers jumped about as much as scores for their fluent fellow students" (Kerr, 2000). Kerr's story noted that test scores had risen for all students in the state about equally, that the Stanford-9 was not written for English learners and is arguably inappropriate, and provided much stronger objections from the research community.

Only three news stories appeared in major U.S. newspapers before the *New York Times* story, one in the *Los Angeles Times* (Groves, 2000, p. A3) and two in the *San Diego Union-Tribune* (Moran & Spielvogel, 2000, p. B1). Like the AP story, these papers presented a much more balanced account. Groves' *Los Angeles Times* story began,

> California students who are not proficient in English improved their scores on the Stanford 9 standardized test at about the same rate as their fluent classmates, but new state data released Monday continue to show an immense disparity between the two groups. (p. A3)

The main *San Diego Union-Tribune* story (Moran & Spielvogel, 2000) opened this way:

> Celebrated gains in student state *test scores* are spread among all students—whether advantaged or disadvantaged, whether they speak English or not—according to data released today. (p. B1)

However, the view appearing in the *New York Times,* due to the paper's enormous influence on the national press, strongly predominated. The *Times* was cited as an authority on the issue in 56 published letters and editorials, and in one story appearing in the *Arizona Republic* (Gonzalez, 2000, p. EX1). Following the appearance of the *Times* article, numerous television and radio news shows, including those of the major television networks, broadcasted the story that rising scores in California indicated Proposition 227 was a success in that state. A story in *Newsday* (Willen & Kowal, 2000, p. A10) said that the conclusion followed from "a recent California study."

Inaccuracies in scientific and technical reporting are known to occur widely in journalistic writing (Simon, Fico, & Lacy, 1989; Singer & Endreny, 1993; Tankard & Ryan, 1974; Weiss & Singer, 1987). However, what is particularly disturbing about the *New York Times* story is that conclusions were drawn based on claims that disregarded basic principles of scientific research design and educational measurement. Undermining the credibility of the story were inadequate consideration of alternative explanations and improper interpretation and use of Stanford-9 scores. Further, the *Times* failed to discuss controlled studies comparing bilingual education to all-English instructional approaches (Ramirez et al., 1991; Willig, 1985) or recent comprehensive research syntheses prepared by the National Research Council (August & Hakuta, 1998; Meyer & Fienberg, 1992). These errors and exclusions are particularly grievous given the high-stakes nature of the inferences drawn regarding an extremely complex educational issue.

Brief Background on Proposition 227 Implementation

In this section, we provide some background on Proposition 227 and summarize briefly recent research addressing the implementation of the initiative. The full text of the California law can be reviewed online (English Language Education for Immigrant Children, 2001; http://www.leginfo.ca.gov/calaw.html), however the general mandate of Proposition 227 is the following: "Children who are English learners shall be educated through sheltered English immersion during a temporary transition period not normally intended to exceed one year" (Section 305, 2001). The law applies to English learners, defined as "a child who does not speak English or whose native language is not English and who is not currently able to perform ordinary classroom work in English, also known as a Limited English Proficiency or LEP child" (Section 306, 2001). The law also further defines sheltered English immersion as "an English language acquisition process for young children in which nearly all classroom instruction is in English but with the curriculum and presentation designed for children who are learning the language" (Section 306, 2001). The implementation of sheltered English immersion (SEI; equivalently referred to as 'structured' English immersion) as mandated by Proposition 227 has been addressed in the educational research literature. Most of these studies may be described as either qualitative studies of the implementation of SEI in California schools or quantitative summaries of standardized achievement scores pre- and post-implementation of Proposition 227.

Studies of Proposition 227 Implementation

Although districts, schools, and teachers did not ignore Proposition 227, there was not a "sea of change" in programs for English learners apparent in the schools (García & Curry-Rodríguez, 2000). In fact, prior to implementation of Proposition 227, only 29% of English learners were in programs that included native language instruction, and 12% of students were still in those programs following implementation (Gándera et al., 2000). Maxwell-Jolly (2000) studied the interpretation and implementation of 227 in seven different school districts and found that although district interpretation of 227 set the tone, responses to and implementation of district policy regarding 227 varied widely. Further research indicated that when district administrators set a strong tone for eliminating native language instruction or providing alternatives to SEI, schools followed suit. However, when district leadership was lacking, implementation of the proposition varied across schools (Gándera et al., 2000).

Gándara (2000) documented the impact 227 had on instructional services, classroom pedagogy, and distribution of teachers, concluding the greatest impact of Proposition 227 was on classroom instruction. For example, teachers reported leaving out much of their normal literacy instruction, such as storytelling and story sequencing, to focus on

English word recognition. Instructional challenges presented by Proposition 227 included having a lack of instructional materials and teaching students with a wider linguistic range (Schirling, Contreras, & Ayala, 2000). Teachers reported that even for programs in which parental waivers were obtained for native language instruction, they were required to include 30 days of English instruction before the waiver could take effect. Because schools did not know how many waivers they would receive, orders for instructional materials were delayed or made in insufficient quantities.

Hayes & Salazar (2001) evaluated instructional services offered to English learners enrolled in SEI in first, second, and third grade classes in Los Angeles Unified School District. They noted uneven implementation of SEI, with programs generally adopting one of two general approaches: use of primary language *for clarification* only and use of primary language for *concept development*. The effects of the proposition on teachers varied based on what the teachers had done prior to the passage of 227 and on teachers' education, skills, experience, and views on student learning (Gándera et al., 2000). For example, teachers who were certified to teach bilingual education were more likely to continue some level of native language support in their classrooms.

Studies offering various other perspectives on implementation of Proposition 227 have been published, many of them in a special issue of the *Bilingual Research Journal* devoted to the topic (e.g., Dixon, Green, Yeager, Baker, & Fránquiz, 2000; Palmer & Garcia, 2000; Paredes, 2000; Schirling, Contreras, & Ayala, 2000; Stritikus & Garcia, 2000). A California Research Bureau report by de Cos (1999) presented issues surrounding implementation of 227 in a historical context of language policy issues. The effects of STAR on English learners were also discussed and the author warned against using these publicly available test scores to evaluate SEI programs. We now review some published quantitative analyses of these STAR data that have been used to support arguments for or against Proposition 227.

Analyses of California Achievement Data

When the publicly available aggregated standardized test scores of California children were released following implementation of Proposition 227, they were quickly analyzed in an attempt to determine the effects of the initiative. It is well-documented in the literature that LEP students made gains in test scores, as did all students in the state (Butler, Orr, Gutierrez, & Hakuta, 2000; Gándara, 2000; Garcia & Curry-Rodríguez, 2000). Butler et al. (2000) reported that schools maintaining strong bilingual programs had scores that equaled or exceeded those of schools that had dropped bilingual programs. In addition, Butler et al. (2000) noted that due to regression to the mean, scores of lower performing students are more likely to improve than those at the middle of the scale. Finally, they emphasized there was significant variation in test scores across schools in both the bilingual and English-only categories. García and Curry-Rodríguez (2000) studied a random sample of districts and found no specific patterns of test scores across schools with different 227 implementation strategies.

Amselle and Allison (2000) examined percentile rank increases for LEP students and found that LEP students made "significant gains in reading and writing in English as well as math" (p.1). They went on to examine percentile rank improvements in four school districts that reported to be in strict compliance with Proposition 227 and four school districts that reported maintenance of a bilingual program. They found greater score improvements in the select districts reporting compliance with the initiative. Finally, they pointed to Los Angeles Unified School District as a district that openly defied Proposition 227 and had percentile rank test scores below "the state average for LEP students" (p. 12). Unfortunately, Amselle and Allison (2000) failed to note the variability within districts. In addition, their focus on select districts did not allow them to examine the variability across districts that reported similar implementations of Proposition 227. Finally, they inappropriately used summaries of national percentile ranks to determine academic growth, a problem we discuss in greater depth later in this paper.

A noteworthy limitation of the publicly available data is the lack of student level information (Gándara, 2000). However, Gutiérrez, Asato, and Baquedano-Lopez (2000) acquired and utilized student-level data for LEP students from an urban unified school district. Over three years, they tracked student scores in this predominantly English-only, phonics-based literacy district. They found the percentage of LEP students scoring at or above the 50th percentile decreased dramatically over the three years. Disaggregation of these data by language group showed that the percentage of Spanish-speaking children reading at or above the 50th percentile dropped from 32% in the first grade to 30% in the second grade to 15% in the third grade. Other language groups (Cantonese, Russian, Hmong, and Mien) also experienced sharp declines between first and third grade (Gutiérrez et al., 2000). The specific causes of these declines were not explored.

In sum, published qualitative evaluation reports of Proposition 227 generally conclude the overall effect of the new law on the education of language minority students has been negative. Furthermore, with the exception of Amselle and Allison's (2000) report, quantitative analyses of the Stanford-9 data to date reveal comparable gains for English learners and their fluent English-speaking peers. Many arguments and quantitative summaries based on the STAR data have been replete with improper statistical analyses and fail to acknowledge the many limitations of these highly aggregated standardized achievement data (e.g., Amselle & Allison, 2000). We now discuss multiple validity concerns as they apply to use of the California Stanford-9 data for evaluating language policy.

Of utmost concern when using assessment data in research should be the validity of the assessment for the intended purpose. A large literature exists addressing the conceptualization of validity in educational and psychological testing and research (see Messick, 1989, for a comprehensive discussion of validity); therefore, we do not attempt a comprehensive review of validity, but rather concentrate on those validity issues that appear to be most problematic in our research context. To help focus our discussion, we borrow from Messick (1989) a definition of validity as a unified concept with multiple facets: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis added). As such, validity is not merely about the meaning of test scores. Validity encompasses "... the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (Messick, 1989, p. 13). The measurement context and the inferential context are both vitally important in forming validity judgments. We focus our discussion in the following section on issues that pose major threats to the validity of inferences based on the STAR data concerning LEP students: language of the test, alternative explanations, and sample selection.

Language of the Test

In this section, we consider the meaning of scores in the context of the assessment. Of particular concern is the administration of English-language standardized achievement tests to evaluate the academic achievement of students who are not proficient in English. Testing students in a language in which they are not yet proficient is problematic for multiple reasons. The *Standards for Educational and Psychological Testing* warn that when testing a non-native speaker in English, the test results may not reflect accurately the abilities and competencies being measured if test performance depends on the test takers' knowledge of English (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The Stanford-9 is a test of academic achievement, not a test of language proficiency, and the test developers have not conducted any specific studies to establish validity of Stanford-9 scores for children who have limited ability in the language of the test (Harcourt Brace Educational Measurement, 1997c). Therefore, limited English proficiency should be regarded a likely source of measurement error in the Stanford-9 test scores intended to reflect academic achievement.

Language proficiency in general has been shown to influence performance on achievement tests (Ulibarri, Spencer, & Rivas, 1981). Pilkington, Piersel, and Ponterotto (1988) reported that the home language of a child influenced the predictive validity of kindergarten achievement measures. These studies suggest language proficiency plays a role in young children's performance on achievement tests. This relationship may continue in high school children, where LEP status was shown to be a significant predictor of both language arts and mathematics scores on the California Assessment of Progress, although a poverty measure was a stronger predictor (Wright & Michael, 1989).

To make clear the validity problem with testing LEP students in English, imagine that a test of academic achievement was administered to typical U. S. elementary school students in Spanish. Because relatively few U. S. students know Spanish at that age, they would have considerable difficulty understanding the questions on the test, and consequently would be expected to perform quite poorly. Because the test purports to measure academic achievement and not knowledge of Spanish, the use of Spanish as a medium for the test is a source of error in the measurement of academic achievement in this population of students.

A central problem is that a student who does not know the language of the test may have considerably more knowledge of academic content than the test can detect. Consider, for instance, the following sample item for 3rd-grade mathematics, provided by the Stanford-9 publisher (Harcourt Brace Educational Measurement, 1997a):

> Herb bought a candy bar for $0.75 and a package of gum for $0.50. What else do you need to know to find out how much change Herb should receive?
>
> A. How many sticks of gum were in the package.
> B. Where he bought the candy and gum.
> C. The size of the candy bar.
> D. How much money he gave the clerk.

A 3rd-grade English learner may know the answer to this question, but the test will not detect this if the child's English is so limited that he or she cannot interpret the question. The structure of the question is, in fact, extremely complex, involving multiple cases of what linguists call long-distance *wh*-extraction, posing a special burden on cognitive resources devoted to language processing (Stabler, 1986, 1994) which is known to introduce particular processing difficulty for second language learners (Juffs & Harrington, 1995, 1996; Myles, 1996).

It is also commonly believed the Stanford-9 can be used as a proxy measure for English language development, the rationale being that because children must know English reasonably well to understand the questions on the test, lower scores reflect limited knowledge of English and higher scores reflect greater knowledge of English. However, children might score low on the Stanford-9 for reasons having nothing to do with knowledge of English, so we cannot infer that LEP children who have lower scores on the test have not learned English. Other factors, such as socio-economic status and educational background, have a documented relationship with children's academic growth in

89

school (Berliner & Biddle, 1995; Genesee, 1984; Rosenthal, Milne, Ellman, Ginsburg & Baker, 1983). Thus, children may know English reasonably well and still do poorly on the Stanford-9 as a result of other factors.

It may also appear plausible to assume rising test scores can be interpreted as an indication that LEP children are improving in English. However, the Stanford-9 items were not developed in relation to a theory of language proficiency. Consider, for instance, the following sample test item, provided by the Stanford-9 publisher (Harcourt Brace Educational Measurement, 1997a), which targets 3rd-grade language arts. The test taker is asked to select the corrected version of the prompt sentence.

> Sara walked into the kitchen, she looked for a snack.

> A. Sara walked into the kitchen. And looked for a snack.
> B. Sara walking into the kitchen and looking for a snack.
> C. Sara walked into the kitchen and looked for a snack.
> D. Correct as is.

This item tests knowledge of writing conventions related to the use of commas, a domain of academic achievement. A Spanish-speaking child familiar with basic writing conventions in Spanish could answer this question correctly with extremely limited knowledge of English grammar and vocabulary. Apart from knowing about comma placement in Spanish, the child only needs to know the English words *and*, *walk* and *look* (or needs to infer that *walk* and *look* are verbs from implicit or explicit knowledge of their morphology, *-ed* and *-ing*), and to have a tacit sense that *into the kitchen* and *for a snack* have some relation to the verbs. Choice C will be immediately favored in light of what the child knows about comma placement in Spanish. Choices A and B can be ruled out with certainty if the child also knows that progressives require a helping verb, as they do in Spanish, and that English sentences require overt subjects. In other words, a child who knows the relevant academic concept could get this item right with very little knowledge of English.

Inferences made from Stanford-9 achievement test scores can only be made with respect to the intended content domain, and as we have discussed, even inferences to the content domain are of questionable validity for LEP students. August and Hakuta (1989) suggested there is a need to develop guidelines for determining the proficiency levels at which English learners are ready to take the same assessments as their English speaking peers (see also Duran, 1989, for a review of language proficiency assessment as well as other testing issues for linguistic minorities). Further, Messick (1989) argued convincingly that social and educational consequences of measurement-based inferences are a vital part of the validity framework. Social and educational consequences for LEP students may include program placement, implications for self-concept, funding, grade level retention/promotion, teacher bonuses, and school evaluation. Using aggregated Stanford-9 scores for evaluating language programs and informing policy decisions may have dramatic social and educational consequences; therefore threats to validity relating to the language of the test should be carefully considered.

Alternative Explanations

An important consideration in interpreting trajectories of achievement scores in California is the acknowledgement of potential confounding conditions and alternative explanations. In this section, we address five such considerations: simultaneous changes in educational policy and practice, inconsistent implementation of immersion programs, increasing test familiarity and preparation, the limitations of using aggregated data, and regression to the mean.

Simultaneous Policy Implementations

Proposition 227 was introduced concurrently with other changes in educational policy and practice. In fact, Gándara (2000) explained Proposition 227 was enacted in what has been the most active period of education reform in recent times. Statewide initiatives include class size reductions from an average of 30 to 20 in early elementary classrooms and a switch from a whole language approach to a phonics-based method of reading instruction for poor readers. Gutiérrez et al. (2000) specifically noted that class size reduction, the new state standardized testing program, new reading and accountability initiatives, and the new language arts standards had all been implemented concurrently. In addition, other reforms have likely occurred at the district, school, and classroom level. Any or all of these may be important contributors to student gains.

Inconsistencies in Language Programs

Evaluating language program policy is further complicated by inconsistent implementation of English immersion programs, with instructional practices varying widely across districts, schools, and even classrooms (Berliner, 1988; Gándara et al., 2000). There are many different variations of English-only programs, as well as of bilingual programs. Therefore, it is not clear exactly which programs are being compared when simply examining changes in test scores from 1998 to 1999. The state education system is terrifically diverse and educational practices are far from uniform.

Test Familiarity and Coaching

Because the aggregated scores are publicly released, schools and districts feel pressure to achieve high test scores and thus encourage test preparation in varying degrees. As the stakes become higher, test preparation often becomes a high profit industry, as documented in Texas (McNeil, 2000; Sacks, 1999). The California test was tied to different, but motivating, rewards and sanctions for schools, teachers, and students. Schools and teachers could receive bonuses based on increased test scores. For example, each of the 1,000 certificated staff in underachieving schools with the largest growth in California receives $25,000 (Public Schools Accountability Act, 1999). On the other hand, schools that do not meet goals for academic improvement in 24 months may be taken over by the state Superintendent for Public Instruction. The Superintendent may then take a variety of actions, up to and including closing the school (Public Schools Accountability Act, 1999). Bilingual and English-only teachers alike, in the presence of so many rewards and sanctions, may feel pressure to specifically teach to the test or focus disproportionately on test preparation, as Moran (2000) has discussed.

Even without these extensive consequences, a dramatic and consistent rise in test scores is frequently observed in the first few years following implementation of a new testing program (Linn, Graue, & Sanders, 1990), such as occurred in California following implementation of Stanford-9 testing in 1998. There are several possible explanations for this trend. Coaching and teaching to the test, often at the expense of more desirable teaching and learning activities, can contribute to striking rises in test scores. A meta-analysis of 30 studies revealed that coaching for standardized tests increases test scores in the typical study by .25 standard deviations (Bangert-Drowns, Kulik, & Kulik, 1983). Coaching may refer to a variety of test preparation activities, including general test-taking strategies (e.g., guessing, underlining main ideas, time management), test-specific strategies (e.g., methods useful for quirks of a particular test), and academic instruction tied closely to the content and skills on the assessment (Anastasi, 1981; Bond, 1989).

As teachers and administrators become more familiar with the tests, coaching strategies may become more effective. Butler et al. (2000) suggested that since there is a trend for test scores to rise for all students in California, these broad patterns of improvement may result largely from "teaching to the test." Teachers in all types of classrooms, including bilingual and SEI classrooms, report having modified their teaching practices substantially, with a greater emphasis on preparing students to answer English standardized test-like questions (Alamillo & Viramontes, 2000; Gándara 2000).

Unknown Student and School Characteristics

Another factor limiting conclusions based on the STAR data is the nature of the data itself. We have noted the California STAR data are available to the public and the research community only in aggregated form—by grade level within school. The lack of student-level information makes these data insufficient for thoroughly exploring relations between student achievement and language program for LEP students. Although indicators of the dominant language program and enrollment numbers are available at the school level (California Language Census Data Files, 2000), this information cannot be tied to individual students or even grade-level averages. Further, relevant student-level information such as socioeconomic status, level of English proficiency, and the previous year's score cannot be used to control for potentially relevant individual differences.

A problem of particular relevance for studying the effects of language programs on LEP students is the variability among districts in the criteria for defining LEP students, as well as who will be tested. As noted by Gándara (2000) and Butler et al. (2000), redesignation of students with borderline English proficiency could have a profound effect on aggregated scores in LEP and non-LEP groups. Although the Stanford-9 is not a measure of English language proficiency, some districts redesignate LEP students achieving a certain score on the Stanford-9 to EP status for the following year. This skimming effect may result in depressed scores for the LEP group. In contrast, Gándara (2000) pointed out that some districts are not reclassifying students on the basis of high scores on the Stanford-9, perhaps because they have not performed well on language proficiency tests. Differences in redesignation policies and rates, in the absence of student-level data, blur the meaning ascribed to LEP and EP score means.

Additionally, use of these aggregated data induces two distinct problems related to school size. First, grade-level averages of achievement scores for each academic subject were included only if there were at least ten students in the summary category represented. Schools were therefore unable to report summaries of any subgroup, such as LEP students, if there were fewer than ten students in the grade. It follows that the scores of many students are not represented in these subgroup aggregates. If, for example, schools with fewer LEP students tended to be schools having higher socioeconomic status (SES), systematic omission of these schools due to insufficient numbers of LEP students may introduce bias in estimated LEP group means related to average school SES.

A second problem related to school size is that the oft-reported statewide averages of the grade-level means do not account for the drastically varying numbers of students represented by the available grade-level within-school score aggregates. Certainly, the numbers of students in each grade varies across schools overall and for particular subgroups, yet computation of unweighted averages gives equal influence to small and large schools. This presents a unit of analysis problem when trying to make inferences about achievement at the student level. Unweighted averages of the grade-level means do not provide appropriate estimates of the statewide student averages.

As we later address, weighted means might provide better estimates of student scores. However, even weighted means do not allow representation of students excluded from subgroup summaries resulting from too few students in a category. Further, they do not address problems associated with the lack of relevant student covariates or

91

redesignation of language proficiency status. Using aggregated rather than student-level data severely limits the nature and strength of generalizations that can be made based on these data. It is imperative that researchers realize the implicit limitations and fallacies associated with using such grossly aggregated data.

Regression to the Mean

Another explanation for score gains we must briefly consider is regression to the mean, a topic discussed thoroughly in many classic statistical textbooks (e.g., Campbell & Stanley, 1966; Glass & Hopkins; 1996) but often ignored by researchers as a genuine threat to validity. Regression to the mean refers to the tendency for student scores that are extreme upon initial testing (relative to the overall mean) to drift toward the population mean upon subsequent testing. Regression to the mean is particularly important when gains of extreme groups are of interest—such as in the comparison of low-scoring LEP students to other groups. For all available years of Stanford-9 score reports, LEP student scores are markedly lower than those for non-LEP students. In a district such as Oceanside, whose mean scores for LEP students were extremely low in 1998, scores might be expected to rise upon retesting—even without intervention—due to regression to the mean. Butler et al. (2000) compared low-scoring schools with mostly non-LEP students to low-scoring schools with mostly LEP students and demonstrated that schools with both compositions increased similarly from 1998-2000.

Failure to acknowledge or correctly address regression to the mean has been an issue in other large-scale policy analyses. For example, Camilli and Bulkley (2001), in a critique of Greene's (2001) evaluation of Florida's A-Plus accountability system, argued convincingly for policy analysts to be aware of regression to the mean and to use statistical models that take regression to the mean into account. They noted such approaches have been recently employed in North Carolina's development of growth standards for the state. A detailed discussion of regression artifacts, particularly regression to the mean, may be found in a recent book devoted to this subject by Campbell and Kenny (1999).

Sample Selection

Many, if not most, of the published reports of analyses of California's Stanford-9 scores have been based on the consideration of a small sample of schools. While it is perhaps more feasible to focus on a few schools when attempting a descriptive study of the policy implementation process, it is dangerous to make inferences based on quantitative differences in mean achievement across a small number of select schools. The presence of school and classroom effects on student achievement is well documented. Students sharing a common classroom and/or school environment tend to perform more similarly on achievement tests than students sampled from multiple sites (Muthén, 1991; Thompson, 2000). Demographic similarities, as well as collective experiences of students sharing an educational environment, contribute to these classroom and school effects.

Oceanside initially became the district held up as representative of schools that strictly implemented Proposition 227. Beginning with the emphasis on Oceanside's Stanford-9 gains in press releases from Ron Unz and English-only supporters (e.g., English for the Children, 2000), score gains in this district have been repeatedly cited as evidence of the success of the proposition (Amselle & Adams, 2000; Steinberg, 2000). In response to these claims of success due to SEI, opponents of Proposition 227 pointed out marked gains seen in specific districts maintaining bilingual education. For example, Butler et al. (2000) chose schools nominated by Californians Together, a bilingual advocacy group, for analysis (e.g., Fresno Unified School District; Californians Together, 2000, August 21). They compared these to English-only districts held up by advocates of Proposition 227 as the most successful English-immersion schools.

Schools that maintained bilingual programs likely had administrators and teachers who were highly committed to these programs, given the effort needed to obtain parental waivers for participating children. This degree of commitment to bilingual programs may also suggest exceptionally strong and effective programs. Similarly, it can be argued districts such as Oceanside had atypically strong SEI programs. While comparing what seem to be the most successful districts of each program type is informative, the results of such a comparison should not be used to suggest that the same outcomes would be observed in districts with different characteristics. A characteristic unique to a school or district may contribute substantially to a rise in test scores; this characteristic may or may not be related to the language program. We should not be surprised to see contradictory results and inferences regarding program effects from studies that employ selective sampling of a few specific schools and districts. While comparisons of select schools and districts are informative, we urge caution in making generalizations based on such samples.

Data Analysis Decisions

The remaining issues we address are data analytic problems in summarizing the STAR data and using these summaries to support inferences. Our focus here is on analyses that manipulate students' scores in manners incongruent with the intended purpose of the assessment, and therefore these data analysis problems should also be considered threats to the validity of judgments based on STAR data. First, we discuss the misinterpretation and misuse of scores reported in the form of percentile ranks. Problems in using percentile ranks as a basis for longitudinal inferences result from incongruent norm group compositions, unequal score intervals, and difficulties in computing gains. We then discuss unit of analysis issues associated with using aggregated data to make inferences at

the student level.

Misinterpretation and Misuse of Percentile Ranks

Individual student reports of performance on standardized achievement tests, including the Stanford-9, frequently feature percentile ranks. National percentile rank (NPR) scores indicate percentile ranks for a subtest relative to the national within-grade norm group. The NPR scores reported for students taking the Stanford-9 are derived from distributions of scaled scores broken down by grade and subject (Harcourt Brace Educational Measurement, 1997c). For example, a 2nd-grade student estimated to be at the 56th percentile on the math test would score higher than 56% of the students in the 2nd-grade norm group. Similarly, he or she would score lower than 44% of the students in the 2nd-grade norm group. The popularity of percentile rank scores is likely due to the ease with which they are understood at a practical level—parents are comfortable with the notion of a percentile scale on which their child's relative standing can be located. However, there often are hidden validity problems in utilizing a "relative" comparison group in interpreting achievement scores.

Consider a statement from the earlier-mentioned *New York Times* article describing increasing Stanford-9 achievement scores of LEP students in California: "In second grade, ... average score in reading of a student classified as limited in English increased 9 percentage points over the last two years, to the 28th percentile from the 19th percentile in national rankings, according to the state" (Steinberg, 2000, p. 1A). While this statement may seem quite clear on the surface, there are multiple assumptions about the meaning and comparison of percentile ranks that may cloud the perception of true academic growth. We briefly develop several points, well-known to psychometricians, that discourage the use of percentile ranks as measures for assessing academic gains for a collective group of students.

NPRs Are Norm-referenced Scores

For any one student, percentile ranks indicate only relative standing within a norm group. It follows that NPR scores should always be interpreted with the characteristics of the norm group in mind. The norm sample for the Stanford-9 was balanced to generally represent the U.S. population according to socioeconomic status, ethnicity, and urbanicity, with nonpublic schools oversampled to facilitate a separate norm group. Sampled schools were asked to test students who would typically be tested with other students in regular education classrooms, except those classified as trainable mentally handicapped or severely/profoundly mentally handicapped.

Individual districts and schools were therefore able to include or exclude LEP students and some classifications of special education students according to local policy. The tested student population in California contains a much greater proportion of LEP students than does the Stanford-9 norm group. Specifically, the Stanford-9 spring norm sample contained only 1.8% LEP students (Harcourt Brace Educational Measurement, 1997b). In contrast, California estimates approximately 25% of its students are LEP (Macías et al., 1998). The incongruence between the makeup of the reference group and California with respect to LEP students calls into question the validity of generalizations based on NPR scores for LEP students.

Due to their normative nature, NPRs are not a measure of academic achievement as defined by a level of knowledge or skill. It is possible that a true academic gain may appear as a decline according to the change in NPR across years. For example, a student could display greater mastery than in the previous year, but have a lower percentile rank if students in the norm group scored proportionally higher than the tested student in the second year. It again follows that changes in percentile ranks across multiple years are not well suited for demonstrating improvements in academic knowledge or skill.

NPR Score Increments Represent Unequal Achievement Intervals

To understand more thoroughly the pitfalls of manipulating NPR scores, we consider how NPR scores are derived from the students' raw scores. A raw score is simply the total number of items a student answers correctly on a test. The test publisher determines NPRs through a two-step score conversion process (Harcourt Brace Educational Measurement, 1997b). On a specific test, such as the Stanford-9 5th-grade reading test, the original raw scores from the norm group are first transformed into *scaled scores* by applying item response theory (IRT). The IRT model employed for the Stanford-9 takes item difficulties into account to estimate a proficiency level, or scaled score, that is both independent of the specific items to which the student responds (i.e., the form and level of the subtest may vary) and independent of the group of students to whom the test is administered. These scaled scores are on a single scale for a subject area, so they can be compared across different test forms and grade levels. Scaled scores also have the convenient property of an equal-interval scale that supports comparisons of proficiency level across time for a specific subject test (i.e., a one-unit increase from 1998-1999 on a subject test represents the same amount of achievement growth as a one-unit increase from 1999-2000, regardless of grade level).

To convert scaled scores into NPRs, the cumulative distribution of scaled scores from the norm group is transformed into a roughly uniform distribution of percentile ranks ranging from 1 to 99. When a new group of students is administered the exam, their raw scores are first converted to scaled scores. NPRs are then determined for students in the testing group such that the percentile rank for a specific scaled score reflects the percentage of the norm group scoring at or below that level. Table 2 illustrates conversions among raw, scaled, and percentile rank scores on the

93

5th-grade reading subtest of the Stanford-9, using the spring national norm sample (Intermediate 2 Reading Scores, Form T; Harcourt Brace, 1997b). This reading subtest had 84 items. Because there are 4 choices, it is worth noting that we might expect merely guessing on the exam to yield correct answers to 21 of the 84 items, which would place the score at the 3rd percentile on the NPR scale.

Table 2
Conversions of Select Raw, Scaled, and National Percentile
Rank Scores for the Stanford-9 5th-grade Reading Subtest

Raw score	Scaled scores (Form T)	National percentile rank
80	742	99
75	710	93
70	691	82
65	676	72
60	663	59
55	652	48
50	642	38
45	632	29
40	622	22
35	612	15
30	599	8
25	591	6
20	579	3
15	564	1

(Note: Adapted from Harcourt Brace Educational Measurement, 1997b. Intermediate 2 (grade 5) Total Reading, Form T, spring norm sample.)

This transformation of the distribution of scaled scores into percentile ranks is nonlinear, resulting in a loss of the equal-interval scale property of scaled scores. Equal differences in percentile ranks do not reflect equal differences in achievement or skill. Because the relative frequency of the original scaled scores is typically greater in the middle range than in the upper and lower ranges, conversion to percentile ranks results in spreading of the mid-range scaled scores and condensing of the upper- and lower-range scaled scores (the tails of the distribution).

This lack of an equal-interval scale has several important implications. An achievement gain of 1 scaled score point does not result in a consistent gain in percentile ranks throughout the range of the scale. Specifically, a gain of 1 percentile rank will reflect a greater achievement difference (reflected by the scaled score difference) in the upper or lower range than in the middle range. A difference in percentile ranks between 10 and 20 or between 80 and 90 may reflect a greater achievement gain than a difference between 50 and 60. This can be seen from the example in Table 3. Student 1 scored lower than most other students taking the test and Student 2 scored near the middle of students taking the test. Both students improved 50 scaled score points from 3rd grade to 4th grade, representing equivalent achievement gains. Student 1, at the low end of the scale, improved from the 1st percentile to the 7th percentile, an increase of 6 percentiles. However, Student 2, in the middle of the scale, improved from the 41st to the 67th percentile, an increase of 26 percentiles. In addition, the accuracy of percentile ranks differs across the range of scores (Rogosa, 1999). Comparisons of percentile gains for students at different skill levels are not transparent and may be regarded as misleading at best.

Table 3
Example of Percentile Rank Gains at
Different Points in the Distribution

	Student 1		Student 2	
	Grade 3	Grade 4	Grade 3	Grade 4
Scaled score	525	575	605	655

94

Percentile rank	1	7	41	67

(Note: Adapted from Harcourt Brace Educational Measurement, 1997b. Primary 3 (grade 3) &
Intermediate 1 (grade 4) Total Reading, Form T, spring norm sample.)
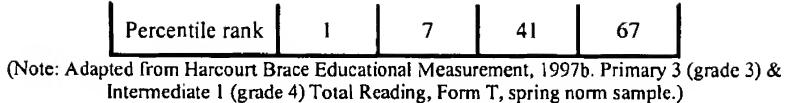
NPRs Should Not Be Averaged or Used to Compute Gains

Another important implication of the lack of an equal-interval scale is that means and gain scores of NPRs should not be computed (Crocker & Algina, 1986; Cronbach, 1960). The California Stanford-9 data are only publicly available as means for each grade level within school. Additionally, these grade-level within-school means often are averaged further in an attempt to summarize subgroup means. Due to the unequal intervals created in deriving the NPR scale, such averaging of data may drastically propagate errors in estimating true achievement.

Although both the California STAR website (California Standardized Testing and Reporting, 2000) and Stanford-9 Technical Manual (Harcourt Brace Educational Measurement, 1997c) state explicitly that NPRs should not be used to determine true academic change across years, many reports have focused on changes in percentile ranks (e.g. Amselle & Allison, 2000; Butler et al., 2000; English for the Children, 8/14/2000; Steinberg, 2000, p. 1A). The major argument against using changes in NPRs to assess achievement gain for a student is that a gain of 1 scaled score point translates into different NPR intervals at different points on the scale, thereby obscuring measurement of true achievement gains. Further, because scores within each year have already been averaged, the collective gains of a group are even more problematic to assess and compare.

Gains have been computed from two perspectives: within-grade changes across years or cohort gains across years. *Within-grade changes*, which have been more commonly reported in the context of California STAR data, compare means for a grade level across subsequent years (e.g., 2nd-graders in 1998 to 2nd-graders in 1999). Even when using scaled scores, these within-grade changes are not true achievement gains in the sense that they are based on different groups of students. In contrast, *cohort gains* compare scores for a cohort of students across subsequent years (e.g., 2nd-graders in 1998 to 3rd-graders in 1999). However, we caution that when working with aggregated data, individual students cannot be tracked and therefore student mobility introduces some uncertainty to within-school cohort gains. Further, even if we assume the student group to be relatively consistent across years, the norm group used to determine the NPRs is different. As earlier described, it is therefore possible for slight gains in true achievement to appear as declines if the relative standing in the norm group is lower in the second year of testing, and vice versa.

In summary, the use of NPRs for tracking and comparing student achievement trajectories is problematic from multiple perspectives. The utility of percentile ranks is limited to informing judgments of how well a student does relative to the norm group for a subject and grade, gaining a view of a student's relative score profile across subjects, and evaluating whether a student has improved standing relative to the norm group from one year to the next. Such judgments are only valid if the norm group is an appropriate basis for comparison, which it quite clearly is not for LEP students. The characteristics of percentile ranks—including norm group inconsistencies and an unequal interval scale—make this score form unsuitable for large-scale longitudinal policy analysis. We now treat one additional data analysis problem we have observed in reports of California achievement results—inappropriate averaging of data.

Averaging Scores Across Subjects and Grades

Regardless of the score form reported, it is incorrect to average scores across different subjects and different grades (California Standardized Testing and Reporting, 2000; Harcourt Brace Educational Measurement, 1997c). Yet we have observed multiple citations of score improvements that involve averages across both grades and subjects. For example, consider the following statement, from a press release on the *English for the Children* website, that attempts to summarize the academic progress of English learners in California:

> From 1998 to 2000, California English learners in elementary grades (2-6) ... raised their mean
> percentile scores by 35% in reading, 43% in mathematics, 32% in language, and 44% in spelling,
> with an average increase of 39% across all subjects (English for the Children, 2000, August 14).

The same site also displays tables showing mean percentile ranks across elementary grades 2-6 and across multiple subjects. Even prior to computing percentage improvements, consider the layers of averages implied in these numbers:

1. LEP student scores are first aggregated to grade-level, within-school means (before release of data);
2. LEP grade-level, within-school means are averaged across grades (2 through 6) *and* schools;
3. LEP grade-level, within-school means are averaged across subjects (reading, mathematics, language, and spelling) *and* schools; and
4. LEP grade-level, within-school means are then simultaneously averaged across grades *and* subjects *and* schools.

We therefore see that "average increase of 39% across all subjects" relies on a notion of gains based on *means of*

means of means of means. What do these *mean?*

These overall averages are not meaningful or defensible from a measurement perspective and, further, they may obscure important differences in means that exist across grades and subjects. Such data summaries are psychometric nightmares, and are particularly haunting when used to support arguments for educational policy that may strongly impact students' educational opportunities.

An Analysis of the California STAR Data

Motivated by the validity problems we have observed in other summaries of these data, we conducted a reanalysis of the California STAR data. Although we have argued that the Stanford-9 has significant limitations as a measure of achievement for LEP students and that these aggregated data lack information necessary to inform language program policy, it is apparent from our review of press and research reports that trends observed in these data will continue to be cited as evidence for arguments on both sides of the language policy spectrum. Here we attempt to analyze differences in score means and trends for California LEP and EP students and interpret them thoughtfully—without leaping to unwarranted inferences about language program effects. We present a comprehensive summary of means and gains for all grades and subjects tested; however, we focus our discussion on reading, language (Note 2), and mathematics scores for the elementary grades 2-6.

Methods

Data

We compiled data from three publicly available data sources. First, Stanford-9 scores were obtained from the California STAR website (California Standardized Testing and Reporting, 2000). This dataset provided within-school grade-level means on subtests of the Stanford-9 for reading, mathematics, and language for grades 2 to 11, spelling for grades 2 through 8, and science and social studies for grades 9 through 11. In addition to the within-school grade-level means, STAR also reports subgroup means for EP and LEP students. However, as noted previously, data were not reported for groups of less than 10 students for reasons of confidentiality. For example, the grade-level aggregate scores for the LEP subgroup were not included in the data report if a grade had less than 10 LEP students. Additionally, we obtained supplemental demographic information from the language census website (California Language Census Data Files, 2000) and from the academic performance index data website (California Academic Performance Index Data Files, 2000).

Statistical Procedures

The outcome scores used in these analyses were in the form of subject-area scaled scores. Recall that scaled scores are academic proficiency estimates that can be compared across time and across different levels of a subject-area test. Weighted means were computed for each subject area and grade level for three groups. With weighted means, schools with more students are weighted more heavily than schools with fewer students in computing the overall mean, providing a closer approximation to student-level mean. The three groups for which means were computed were: all students, LEP students, and EP students. Schools reporting overall scores were used in computing weighted means for the group of all students. For LEP and EP subgroup means, we included all schools that reported subgroup means for *both* LEP and EP students. In 1998, however, the STAR dataset did not include aggregate scores for EP students separately, so we were unable to compare these groups in 1998.

In order to determine changes in scores across years, we computed both within-grade changes and cohort gains. First, within-grade changes were computed by subtracting within-school grade-level means from one year to the next for a single grade; an example is the difference between 4th-grade reading scores in 1998 from 4th-grade reading scores in 1999. The weighted means of these within-grade changes were then computed for each grade level in each subject area. Second, cohort gains were computed by subtracting within-school grade-level scores from one year to the next in consecutive grades; an example is subtracting 3rd-grade reading scores in 1998 from 4th-grade reading scores in 1999. We regard this as a *loose* cohort because we do not have evidence regarding which students remained in the same school from one year to the next. Again, the weighted mean of these gains was computed in each subject area.

Results

Descriptive Statistics

In order to examine how grade level means change for each academic subject over the years 1998, 1999, and 2000, we computed weighted means and standard deviations for each grade in each subject (see Appendix A). Our main finding from these means is that over the three-year period, scores for LEP students remain substantially below the scores for EP students in schools that reported aggregate scores for both LEP and EP students. And, with few exceptions, the gap in LEP and EP students' scores does not appear to be narrowing. We describe the score trends for 2nd through 6th grades in reading, mathematics, and language in greater depth in the following section. We summarize mean score differences by examining within-grade changes, followed by cohort gains.

96

Within-grade changes

Reading. Within-grade changes for each grade in reading across consecutive years are shown in Table 4 for grades 2-6 and in Appendix B for all grades. Figure 1 shows that for 2nd grade, all student groups improved substantially from 1998 to 2000. For example, from 1999 to 2000, 2nd-grade LEP students gained an average of 4.20 scaled score points, EP students gained an average of 4.40 scaled score points, and all students gained an average of 5.39 scaled score points. (Recall that the group termed *all* students consists of a larger number of schools; LEP and EP means are based on schools reporting scores in both of these subgroups.) It is also interesting to note that although the overall means increased, gains varied considerably among schools, and not all schools experienced improvement. For example, from 1999 to 2000, 27.2% of schools experienced declines in mean second-grade reading for LEP students and 28.6% experienced declines for EP students.

Table 4
Weighted Mean Within-Grade Gains in
Reading for Grades 2-6

Grade		1998-1999		1999-2000			1998-2000	
		LEP	ALL	LEP	EP	ALL	LEP	ALL
2	M	7.84	5.88	4.20	4.40	5.39	12.90	11.21
	SD	9.82	8.89	10.54	9.67	8.31	10.88	9.66
3	M	8.30	5.02	3.12	4.72	4.57	11.58	9.60
	SD	10.87	8.36	10.98	9.41	7.80	10.24	8.71
4	M	6.47	2.69	2.16	3.58	3.79	7.60	6.48
	SD	10.27	8.02	10.07	8.69	7.48	9.65	8.44
5	M	3.81	1.65	1.27	1.88	1.83	4.54	3.44
	SD	7.68	7.09	7.92	7.75	6.90	8.88	7.56
6	M	3.56	2.12	1.96	1.97	1.51	4.18	3.68
	SD	8.01	5.93	6.84	6.52	5.52	7.35	6.23

Figure 1. Mean within-grade changes for 2nd-grade reading.

A comparison of reading scores for LEP and EP students from 1999 to 2000 (EP students were not reported separately in 1998) across grades 2-6 indicated that EP students made slightly larger gains than LEP students in all 5 grades. In no instance, however, was the difference in gains more than two scaled score points. This pattern is illustrated in graphs of reading means for grades 1 and 4 (see Figures 1 and 2, respectively), which show nearly parallel lines for LEP, EP, and all students.

98
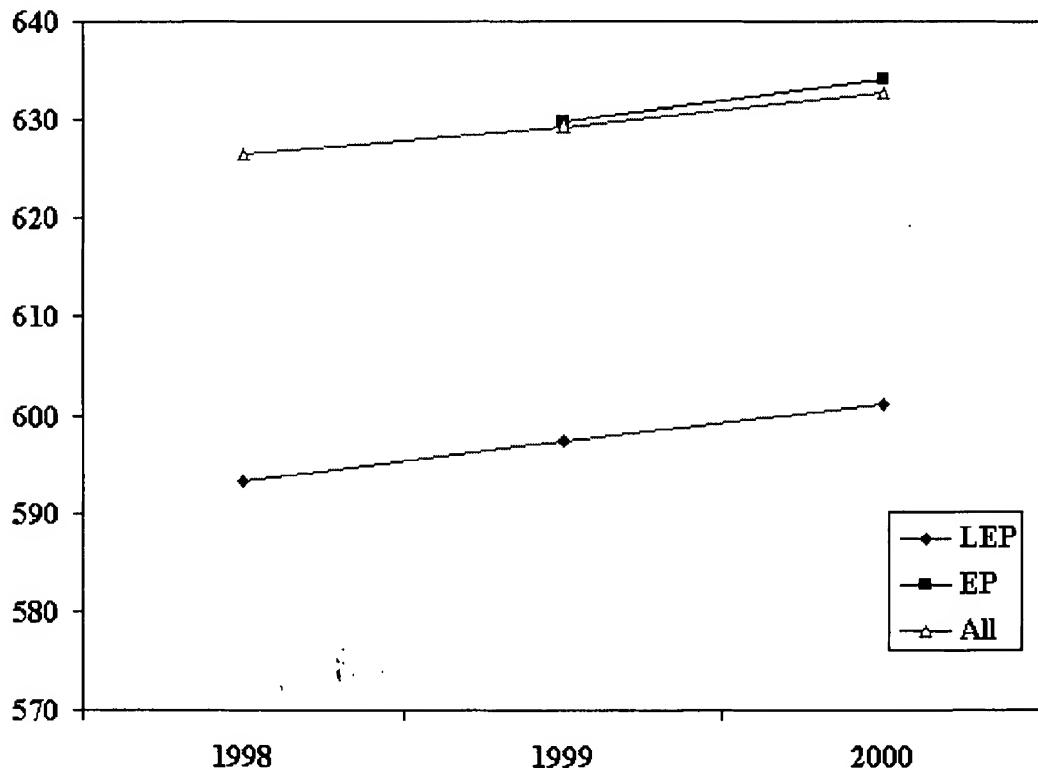
Figure 2. Mean within-grade changes for 4th-grade reading.

Language. Within-grade changes in language were similar to those in reading (see Appendix B). Students in each group displayed gains in scores from 1998 to 1999, 1999 to 2000, and 1998 to 2000. A comparison of LEP and EP students from 1999 to 2000 again revealed that EP students made slightly larger gains than LEP students in grades 2-6, although the overall improvement was never greater than three scaled score points.

Mathematics. An examination of the within-grade mean increases in mathematics scores revealed that LEP students again improved slightly less than EP students in all grades from 1999 to 2000, as reported in Table 5 for grades 2-6 and Appendix B for all grades. For example, in 2nd grade, the mean change for LEP students was 6.91 scaled score points and the mean change for EP students was 7.63 scaled score points. In 4th grade, LEP students had an average increase of 4.95 scaled score points, while EP students had an average increase of 7.33 points. Figures 3 and 4 illustrate within-grade improvements for 2nd and 4th grades, respectively. A visual inspection suggests that increases were similar across groups; however, the cumulative effect of greater improvements for EP students over multiple years makes the trend worth noting. We again caution that these within-grade changes are average score improvements across years based on different groups of students.

Table 5
Weighted Mean Within-Grade Gains in
Mathematics for Grades 2 through 6

		1998-1999		1999-2000			1998-2000	
Grade		LEP	ALL	LEP	EP	ALL	LEP	ALL
2	M	10.08	7.72	6.91	7.63	7.29	14.37	15.04
	SD	12.82	10.02	14.14	12.01	10.18	13.47	11.66
3	M	11.82	8.17	7.35	9.60	8.38	16.43	16.56
	SD	13.23	9.87	12.92	10.59	9.38	13.08	11.03
4	M	7.84	4.88	4.95	7.33	6.91	11.30	11.83
	SD	10.33	8.71	11.39	9.51	8.35	11.08	9.69
5	M	4.90	3.78	4.49	5.68	5.34	8.44	9.09
	SD	9.10	8.31	10.03	8.84	8.18	10.37	9.39
6	M	4.66	4.30	3.67	4.89	4.11	7.55	8.49
	SD	10.06	7.54	9.66	8.61	7.51	9.51	8.51
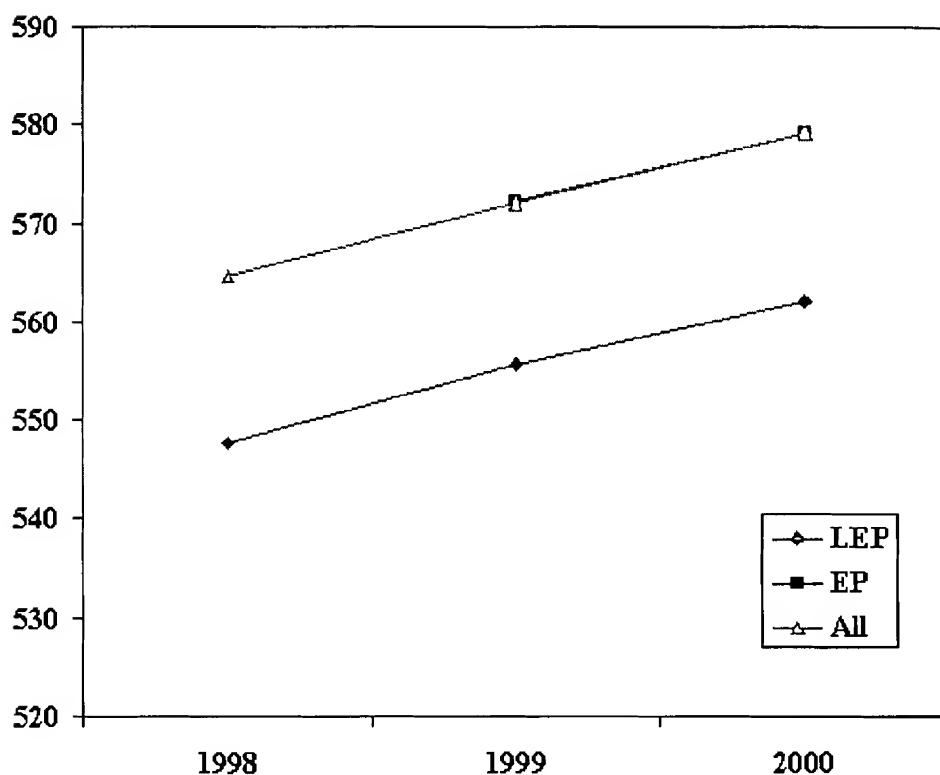
Figure 3. Mean within-grade changes for 2nd-grade mathematics.
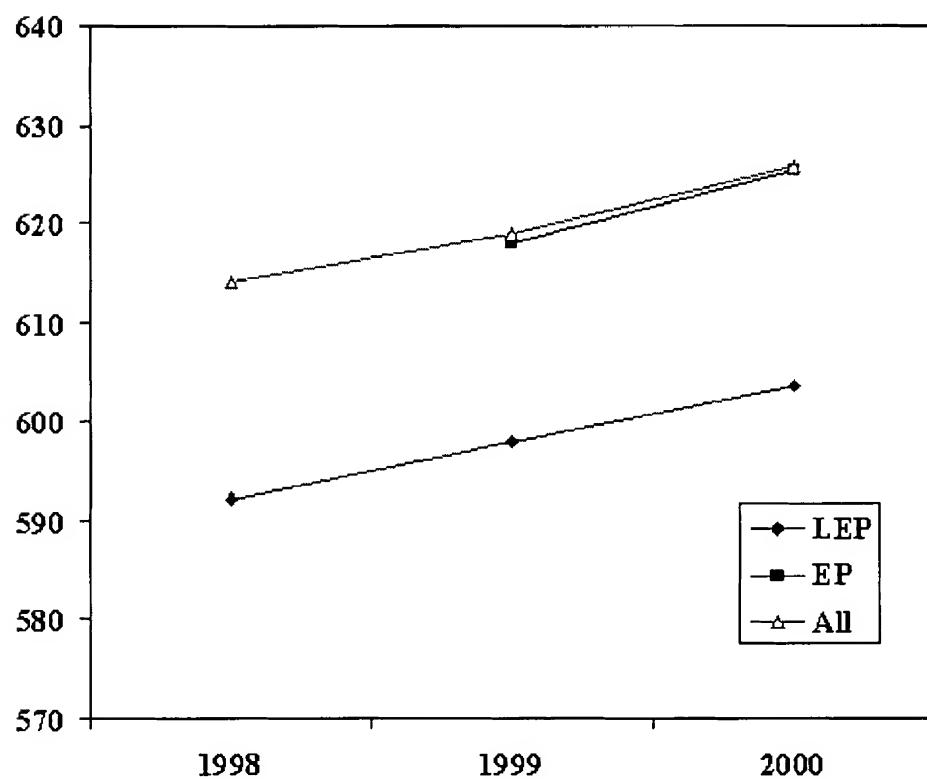
Figure 4. Mean within-grade changes for 4th-grade mathematics.

Cohort gains

100

Reading. We examined the gains made by cohorts of students across the three years of the test (see Table 6). Figures 5 and 6 show cohort gains for grades 2-4 and grades 4-6, respectively, in reading. Both LEP and EP cohorts improved substantially from 1999 to 2000; however, there was not a clear pattern with respect to which group gained more. From 2nd to 3rd grade, LEP students gained less (28.70 scaled score points) than EP students (34.21). The two groups' gains were similar to each other from 3rd to 4th grade and from 4th to 5th grade, while the LEP students gained more than EP students from 5th to 6th grade. Another interesting trend is that for all groups, cohort gains across grades are much greater in early elementary grades (2-4) than upper elementary grades (4-6).

Table 6
Weighted Mean Cohort Gains in Reading for Grades 2 through 6

Grades		1998-1999		1999-2000			Grades	1998-2000	
		LEP	ALL	LEP	EP	ALL		LEP	ALL
2 to 3	M	30.59	34.00	28.7	34.21	32.66	2 to 4	58.33	62.19
	SD	9.51	8.48	9.83	9.36	8.61		11.01	9.20
3 to 4	M	30.27	29.54	27.87	27.76	28.35	3 to 5	47.99	47.05
	SD	8.89	7.28	8.97	7.85	7.11		10.06	8.93
4 to 5	M	19.51	18.31	17.91	16.75	17.53	4 to 6	38.75	33.53
	SD	8.43	7.06	8.35	7.49	6.79		10.94	9.77
5 to 6	M	19.75	15.72	18.88	14.91	15.92			
	SD	7.41	7.24	8.23	7.27	7.00			

Figure 5. Mean cohort gains for 2nd through 4th graders in reading.

101
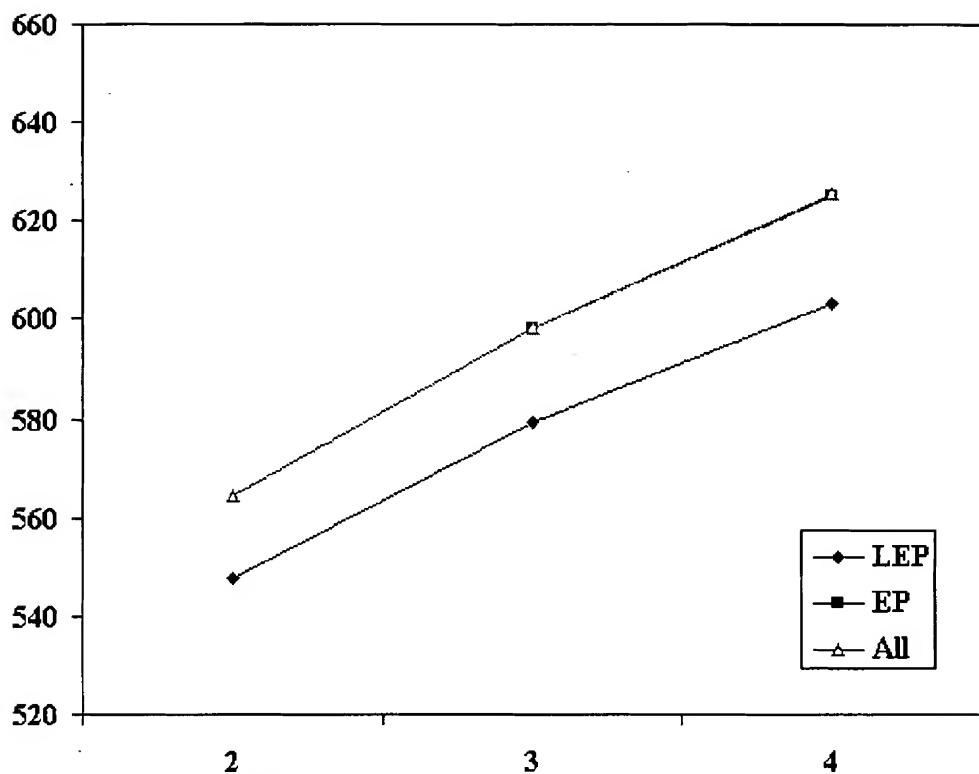
Figure 6. Mean cohort gains for 4th through 6th graders in reading.

Language. The cohort gains in language were similar to those in reading (see Appendix C). There was not a consistent pattern of gains for LEP, EP, and all students. From 1999 to 2000, EP students gained more (22.48 scaled score points) than LEP students (21.02) from 2nd to 3rd grade, although not by much. The two groups gained similarly across the other grade ranges, with EP students gaining slightly more then LEP students from 4th to 5th grade. LEP students gained slightly more than EP students from 3rd to 4th grade and 5th to 6th grade.

Mathematics. The cohort gains for mathematics, displayed in Table 7, reveal a pattern consistent with that seen in the within-group changes. From 1999 to 2000, LEP students gained less than EP students in every cohort. For example, from 2nd to 3rd grade, LEP students gained an average of 31.76 scaled score points, while EP students gained an average of 35.30 scaled score points. These patterns are suggested in Figures 7 and 8 as well, as the LEP line diverges slightly from both the EP and all lines.

Table 7
Weighted Mean Cohort Gains in
Mathematics for Grades 2 through 6

Grades		1998-1999		1999-2000			Grades	1998-2000	
		LEP	ALL	LEP	EP	ALL		LEP	ALL
2 to 3	M	31.61	33.74	31.76	35.30	34.37	2 to 4	55.96	60.70
	SD	12.23	11.03	12.61	12.02	11.20		13.21	11.64
3 to 4	M	26.81	28.30	24.38	27.49	27.07	3 to 5	52.83	57.00
	SD	11.02	9.65	11.12	10.74	9.76		12.85	11.17
4 to 5	M	27.13	28.26	26.07	28.91	28.73	4 to 6	50.46	53.54
	SD	9.43	8.16	9.53	9.03	8.08		12.65	10.95
5 to 6	M	22.79	24.58	22.60	25.67	25.43			
	SD	9.36	8.96	9.64	9.32	8.80			

102

Figure 7. Mean cohort gains for 2nd through 4th graders in mathematics.
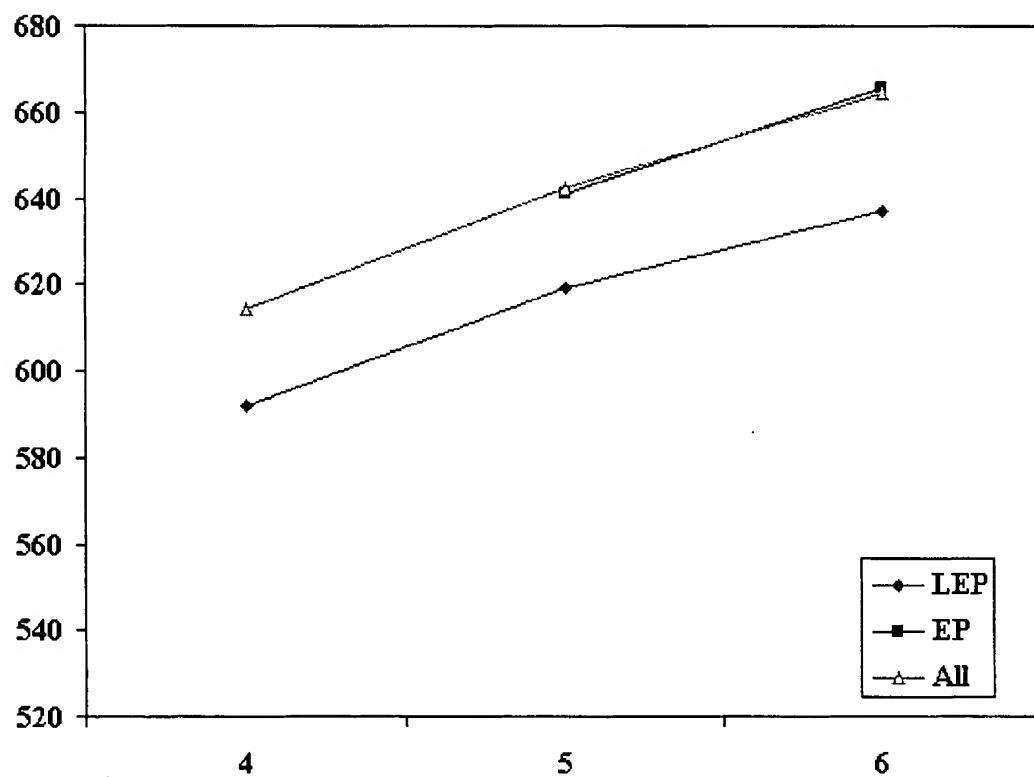
Figure 8. Mean cohort gains for 4th through 6th graders in mathematics.

Comparison of Weighted and Unweighted Means

103

To determine whether the use of weighted means yields substantially different results than unweighted means, we computed a limited number of unweighted means. The unweighted scaled score means for reading, language, and mathematics for grades 2 through 4 are displayed in Appendix D. The results for reading are quite interesting. The unweighted means for the LEP students are higher than the weighted means for all three grades in 1999 and 2000. In contrast, unweighted means for the EP students are lower than the weighted means for all three grades in these years. This indicates that when using unweighted means to summarize reading scores, the gap between EP and LEP students appears to be less than it is when using weighted means to estimate the average score at the student level. In other words, the gap in reading scores between LEP and EP students is wider when taking into account the number of students in each subgroup for a grade-level within a school. This phenomenon is also present in the language scores, but to a lesser extent. For mathematics, the unweighted means were consistently higher for both LEP and EP groups.

Discussion

Our concern for basing educational policy on valid evidence of academic success motivated this commentary and analysis. We first sought to provide a summary and validity critique of writings citing Stanford-9 scores in arguments regarding the success of Proposition 227 in California. Multiple issues have threatened the validity of inferences based on the California data concerning LEP students: testing LEP students in English, failing to consider myriad alternative explanations for score trends, and generalizing from a limited and nonrandom sample of schools. In addition, we have observed errors in quantitatively summarizing this large dataset of standardized scores. In the context of the California data, the misuse and misinterpretation of percentile rank scores, the inappropriate averaging of data across years and grades, and the failure to consider the unit of analysis when using aggregated data are common problems. As validity arguments must include judgments about the appropriateness of interpretations made from test scores (Messick, 1989), this critique should raise many doubts regarding conclusions that have previously been drawn from LEP students' scores on the Stanford-9. In addition, the topics discussed in this paper generalize readily to other applications in which large standardized datasets are cited in educational policy debates.

Our analysis of the STAR dataset differed in three important ways from previous summaries of these data: a) we used scaled scores to assess academic gain across years; b) we computed weighted means to account for the number of students represented by an aggregate score; and c) we were modest with respect to the meaning our results hold for informing language program policy. As previously reported (e.g., Butler et al., 2000) means improved for both LEP and EP students over the three-year period. Our examination of weighted means revealed that from 1998 to 2000, scores for LEP students remained substantially below the scores for EP students in schools that reported aggregate scores for both LEP and EP students, and that with few exceptions this gap is not narrowing.

A within-grade comparison of reading scores for LEP and EP students across grades 2-6 indicated that EP students made slightly larger gains than LEP students in all grades. Loose cohort gains for LEP and EP students in reading were similar; however, for some grade intervals LEP students gained slightly more, while for other intervals EP students gained slightly more. The results for language arts subtest scores were similar to those in reading. In mathematics, an examination of the within-grade mean gain scores revealed that LEP students gained slightly less than EP students in all grades from 1999 to 2000. For 1999-2000 cohort gains, LEP students gained less than EP students in every cohort.

Because it was impossible to follow an individual student's growth across multiple years, the comparison of groups from year to year undoubtedly involved the comparison of different students. This problem was complicated by redesignation of students from LEP status to EP status. Not only was it unclear how many students were redesignated, but redesignation criteria differed across districts. Finally, there are many factors that have been repeatedly shown to influence student achievement. For example, LEP students may differ, on average, from their EP peers with respect to socioeconomic status and mobility, but there was no way to control for such differences using these data.

These findings should be regarded as descriptive summaries of the California STAR data, and we caution that these must be interpreted in light of the substantial limitations of these data for research purposes. *Whatever the score differences between LEP and EP students, judgments of the effects of language program policy on LEP student achievement are not warranted by these data.* To further address the question of performance differences between language programs on a large scale, we attempted to use language census data (California Language Census Data Files, 2000) and academic performance index (API; California Academic Performance Index Data Files, 2000) data to tie schools to specific program types. Most schools reported having students in nearly every program type. Because we could not identify individual students, we could not parse the data from schools into program type. We then attempted to compare schools that reported 100% of their LEP students in bilingual programs in 1998, 1999, and 2000 with schools that reported 100% of their LEP students in English immersion programs in those years; however, this resulted in such a drastic reduction in data that we did not feel quantitative comparisons were warranted (only six schools reported 100% of LEP students in bilingual programs over the three years).

Further investigation is also needed to explore the differences in score trends observed when using unweighted versus weighted means, most notably the underestimation of EP and LEP mean differences with unweighted means. Factors associated with school size may offer meaning to these patterns. We attempted to use API data to investigate the relationship between score trends and mobility, socioeconomic status, and class size. However, due to the aggregated nature of these data, we were only able to reach very general and well-known conclusions, such as that schools with lower average SES tended to have lower test scores.

104

The evaluation of policy outcomes is a high-stakes activity requiring more thoughtful and detailed analyses than computing overall group NPR means. In the context of school accountability systems, Camilli and Bulkley (2001) summarized that tying accountability to single achievement outcomes does not automatically shed light on *why* certain changes were noted. They also argued that appropriate and informative use of statistical models for evaluating policy outcomes requires appreciable technical sophistication. We concur with these notions and find them relevant for our context of evaluating language program effects for LEP students. There is a strong need for research that is well-planned and well-executed that seeks to evaluate language program effects with better controls.

We offer several conclusions based on our validity critique and analysis of the Stanford-9 data. First, the scores of LEP students are not catching up to those of their English-proficient peers in any consistent manner across grades and subjects. Second, the success or failure of programs to remedy the disparity between LEP and EP students should be judged by means other than a single academic achievement test administered in English. The construct of language on an achievement test is qualitatively different from language proficiency as measured on an assessment of English as a second language. Using test scores for any purpose requires that we consider the appropriateness of the scores for the intended use and provide evidence to justify this use. In all assessments, not only should the psychometric validity of the tests be considered, but the potential consequences of the test's use must also be judged. Given the changing demographics of the United States, educators, researchers, and policymakers must join forces to establish policy that will provide maximal opportunity for LEP students to learn.

Notes

[1]We conducted a full-text search of the NEXIS/LEXIS Academic Universe archive of major U.S. newspapers using the search terms "bilingual education, test scores, California." "Major U.S. newspapers" are defined by NEXIS/LEXIS as U.S. newspapers listed in the top 50 in circulation in *Editor & Publisher Year Book*. In a manual inspection of the results, we excluded any publication that did not mention the increase in Stanford-9 test scores in relation to Proposition 227 and also included a *Newsday* article (Willen & Kowal, 2000, p. A10) that refers to the event as "a recent California study."

[2]The "language" subtest of the Stanford-9 measures comprehensive language arts proficiency and is intended for use with English-proficient students; therefore it should not be regarded as an assessment of English language proficiency.

References

Alamillo, L., & Viramontes, C. (2000). Reflections from the classroom: Teacher perspectives on the implementation of Proposition 227. *Bilingual Research Journal, 24*, 155-168.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Amselle, J., & Allison, A. C. (2000, August). Two years of success: An analysis of California test scores after Proposition 227. *READ Abstracts* [On-line]. Available: http://www.ceousa.org/html/227rep.html

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*, 1086-1093.

August, D., & Hakuta, K. (Eds.). (1998). *Educating language-minority children*. Washington, DC: National Academy Press.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*, 571-585.

Berliner, D. C. (1988). Meta-comment: A discussion of critiques of L. M. Dunn's monograph *Bilingual Hispanic Children on the U.S. Mainland*. *Hispanic Journal of Behavioral Sciences, 10*, 273-300.

Berliner, D., & Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's schools*. Reading, MA: Addison-Wesley.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Butler, Y. G., Orr, J. E., Gutierrez, M. B., & Hakuta, K. (2000). Inadequate conclusions from an inadequate assessment: What can SAT-9 scores tell us about the impact of Proposition 227 in California. *Bilingual Research Journal, 24*, 141-154.

California Academic Performance Index Data Files. (2000). Sacramento, CA: California Department of Education. [On-

line database, 8/15/2000] Available: http://api.cde.ca.gov/datafiles.html

California Education Code (2001). *English language education for immigrant children*, Division 1, Part 1, Chapter 3. [On-line]. Available: http://www.leginfo.ca.gov/calaw.html

California Language Census Data Files. (2000). Sacramento, CA: California Department of Education. [On-line database, 8/15/2000] Available: http://www.cde.ca.gov/demographics/files/census.htm

California Standardized Testing and Reporting (2000). Sacramento, CA: California Department of Education. [On-line database, 8/15/2000] Available: http://star.cde.ca.gov

California State Board of Education. (2000). *State Monetary Awards Programs Based on the Academic Performance Index.* [On-line] Available: http://www.cde.ca.gov/ope/ae/pages/certstaffact.html

Californians Together. (2000, August 21). *Schools with large enrollments of English learners and substantial bilingual instruction are effective in teaching English* [On-line]. Available: http://www.bilingualeducation.org/news.htm

Camilli, G., & Bulkley, K. (2001, March 4). Critique of "An evaluation of the Florida A-Plus Accountability and School Choice Program." *Educational Policy Analysis Archives, 9* [On-line journal] Available: http://epaa.asu.edu/

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Campbell, P. R. (1994). *Population projections for states by age, race, and sex: 1993 to 2020* (U.S. Bureau of the Census - Current population reports, pp. 17-23). Washington, DC: Government Printing Office.

Crawford, J. (1999). *Bilingual education: History, politics, theory, and practice.* Los Angeles: Bilingual Education Services.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart, and Wilson.

Cronbach, L. J. (1960). *Essentials of psychological testing (2nd ed.).* New York: Harper.

de Cos, P. (1999). *Educating California's immigrant children: Overview of bilingual education* (CRB-99-010). California Research Bureau.

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Eastin (2000). *Eastin releases additional STAR 2000 test results.* [On-line] Available: http://www.cde.ca.gov/statetests/star/pressrelease2000b.pdf

English for the Children. (2000, August 14). *After two years of Prop. 227 English immersion, a huge rise in California's immigrant scores; Pro-bilingual education districts lag behind.* [On-line]. Available: http://www.yeson227.org/releases.html

Gándera, P. (2000). In the aftermath of the storm: English learners in the post-227 era. *Bilingual Research Journal, 24,* 1-14.

Gándera, P., Maxwell-Jolly, J., Garcia, E., Asato, J., Gutierrez, K., Stritikus, T., & Curry, J. (2000, April). The initial impact of Proposition 227 on the instruction of English learners. [On-line]. Available: http://lmri.ucsb.edu/RESDISS/prop227effects.pdf.

Garcia, E. E., & Curry-Rodriguez, J. E. (2000). The education of Limited English Proficient students in California schools: An assessment of the influence of Proposition 227 in selected districts and schools. *Bilingual Research Journal, 24,* 15-36.

Genesee, F. (1984). On Cummins' theoretical framework. In C. Rivera (Ed.) *Language proficiency and academic achievement.* (pp. 20-27). Clevedon, UK: Multilingual Matters.

Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology (3rd ed.)*. Boston: Allyn & Bacon.

Gonzalez, D. (2000, November 8). Bilingual education gets rebuke from state voters. *Arizona Republic*, p. EX1.

Greene, J. P. (2001). *An evaluation of the Florida A-Plus Accountability and School Choice Program*. New York: The Manhattan Institute.

Groves, M. (2000, August 15). English skills still the key in test scores; Stanford 9: Results show vast divide in achievement between students who have language fluency and those who don't. *Los Angeles Times*, p. A3.

Gutiérrez, K. D., Asato, J., & Baquedano-Lopez, P. (2000). "English for the Children": The new literacy of the old world order, language policy, and educational reform. *Bilingual Research Journal, 24*, 87-112.

Hakuta, K. (2001). Silence from Oceanside and the future of bilingual education. Stanford University manuscript. Available at http://www.stanford.edu/~hakuta/SAT9/Silence%20from%20Oceanside.htm.

Harcourt Brace Educational Measurement. (1997a). *Sample questions for the Stanford Achievement Test, Ninth Edition*. San Antonio, TX: Author. [On-line]. Available: http://www.cde.ca.gov/statetests/star/stanford9.pdf

Harcourt Brace Educational Measurement. (1997b). *Spring norms book*. San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997c). *Technical data report*. San Antonio, TX: Author.

Hayes, K., & Salazar, J. (2001). *Evaluation of the Structured English Immersion Program Final Report: Year I*. Language Acquisition and Literacy, Program Evaluation and Research Branch, Los Angeles Unified School District.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and object asymmetries in wh-extraction. *Studies in Second Language Acquisition*, 17(4), 483-516.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning*, 46(2), 283-326.

Kerr, J. (2000, August 15). English learners show gains. *Associated Press*.

Krashen, S. (1996). *Under attack: The case against bilingual education*. Culver City, CA: Language Education Associates.

Linn, R., Graue, E., & Sanders, N. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice*, 5-14.

Macías, R. F., Nishikawa, S., & Venegas, J. (1998). *Summary report of the survey of the states' limited English proficient students and available educational programs and services, 1996-97*. Washington, DC: National Clearinghouse for Bilingual Education.

Maxwell-Jolly, J. (2000). Factors influencing implementation of mandated policy change: Proposition 227 in seven Northern California school districts. *Bilingual Research Journal, 24*, 37-56.

McNeil, L. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Company.

Meyer, M., & Fienberg, S. (Eds.). (1992). *Assessing evaluation studies: The case of bilingual education strategies*. Washington, DC: National Academy Press.

Moran, C. (2000, July 23). School test gains may be illusory, critics say. *San Diego Union-Tribune*, p. A1.

Moran, C., & Spielvogel, J. (2000, August 15). Students share in test gains; however, the gap between advantaged, disadvantaged has widened, data show. *San Diego Union-Tribune*, p. B1.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational*

Measurement, 28, 338-354.

Myles, F. (1996). The acquisition of interrogatives by English learners of French: The role played by structural distance. *Jyvaskyla Cross Language Studies,* 17, 195-208.

New York Times News Service. (2000, August 20). Spanish-speaking pupils gain after cored move to English; students in California improve strikingly with end of bilingual education. *The Baltimore Sun,* P. 3A.

Pilkington, C. L., Piersel, W. C. & Ponterotto, J. G. (1988). Home language as a predictor of first-grade achievement for Anglo- and Mexican-American children. *Contemporary Educational Psychology, 13,* 1-14.

Public Schools Accountability Act. (1999). [On-line] Available: http://www.cde.ca.gov/psaa/

Ramirez, D., Pasta, D., Yuen, S., Billings, D., & Ramey, D. (1991). *Final report. Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual education programs for language-miority children.* (Vols. 1 & 2). San Mateo, CA: Aguirre International.

Rogosa, D. R. (1999). *How accurate are the STAR national percentile rank scores for individual students? An interpretive guide.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing [On-line]. Available: http://cresst96.cse.ucla.edu/CRESST/Reports/drrguide.html

Rosenthal, A., Milne, A., Ellman, F., Ginsburg, A., & Baker, K. (1983). A comparison of the effects of language background and socioeconomic status on achievement among elementary-school students. In K. Baker & E. de Kanter (Eds.). *Bilingual education: A reappraisal of federal policy.* Lexington, MA: Lexington Books.

Rossell, C., & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English, 30,* 7-74.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it.* Cambridge: Perseus Books.

Schirling, E., Contreras, F., & Ayala, C. (2000). Proposition 227: Tales from the schoolhouse. *Bilingual Research Journal, 24,* 127-140.

Simon, T. F., Fico, F., & Lacy, S. (1989). Covering conflict and controversy: Measuring balance, fairness, defamation. *Journalism Quarterly, 66,* 427-434.

Singer, E. & Endreny, P. (1993). *Reporting on risk: How the mass media portray accidents, diseases, and other hazards.* New York: Russell Sage Foundation.

Stabler, E. P. Jr. (1986). Possible contributing factors in test item difficulty: Research memorandum. Princeton, NJ: Educational Testing Service.

Stabler, E. P., Jr. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing.* New York, NY: Erlbaum.

Steinberg, J. (2000a, August 20). Increase in test scores counters dire forecasts for bilingual ban. *New York Times,* p. A1.

Steinberg, J. (2000b, August 20). English-only classes appear a success; after ending bilingual programs, California schools see boosts in test scores. *Milwaukee Journal Sentinel,* p. 3A.

Steinberg, J. (2000c, August 20). Rising test scores shock critics of English-only law. *The Plain Dealer,* p. 21A.

Tankard, J., & Ryan, M. (1974). News source perceptions of accuracy in science coverage. *Journalism Quarterly, 51,* 219-225, 334.

Thompson, M. S. (April, 2000). *Investigating dependency at the classroom and school levels in educational research: A multilevel approach.* Presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.

Ulibarri, D. M., Spencer, M. L., & Rivas, G. A. (1981). Language proficiency and academic achievement: A study of language proficiency tests and their relationship to school ratings as predictors of academic achievement. *NABE Journal, 5,* 47-80.

United States Department of Education. (1994). *Summary of the bilingual education state educational agency program survey of states' limited English proficient persons and available educational services (1992-1993): Final report.* Arlington, VA: Development Associates.

United States Department of Education. (1997). *1993–94 schools and staffing survey: A profile of policies and practices for limited English proficient students: Screening methods, program support, and teacher training.* Washington, D.C.: U.S. Department of Education. [On-line]. Available: http://nces.ed.gov/pubs/97472.pdf

Weiss, C., & Singer, E. (1987). *Reporting of social science in the national media.* New York: Russell Sage Foundation.

Willen, L., & Kowal, J. (2000, September 9). Study: Bilingual ed falling short. *Newsday*, p. A10.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55,* 269-318.

Wright, C. R., & Michael, W. B. (1989). The relationship of average scores of seniors in 172 southern California high schools on a statewide standardized achievement test to percentages of families receiving federal aid and to percentages of limited-English proficient speakers. *Educational and Psychological Measurement, 49,* 937-944.

About the Authors

Marilyn S. Thompson

Marilyn Thompson is Assistant Professor of Measurement, Statistics, and Methodological Studies in the Division of Psychology in Education at Arizona State University. Her research interests include methodological techniques for large data set analysis, applications of structural equation modeling and multilevel modeling, and policy pertaining to the assessment of English learners. Dr. Thompson is Director of the EDCARE laboratory at ASU, providing measurement, statistics, and evaluation services to the educational community. She previously taught physics and chemistry and was science department chair in an urban high school, giving her a first-hand perspective on the role of classroom assessment. She may be reached at m.thompson@asu.edu.

Kristen E. DiCerbo

Kristen DiCerbo is a Ph.D. student in the School Psychology program in the Division of Psychology in Education, College of Education at Arizona State University. Her research interests include psychoeducational assessment, especially the assessment of minority children, and internalizing problems in children. She is currently a psychology intern in a district serving a large population of English learners. Ms. DiCerbo was formerly Editor of *Current Issues in Education*. She may be reached at kristen.dicerbo@asu.edu.

Kate Mahoney

Kate Mahoney is a Ph.D. student in Curriculum and Instruction with a concentration in bilingual education at Arizona State University. Her teaching experience includes teaching mathematics with English language learners. Her research interests include psychometrics and validity issues concerned with testing English language learners.

Jeff MacSwan

Jeff MacSwan is an Assistant Professor of Language and Literacy in the College of Education at Arizona State University. He has published articles in *Hispanic Journal of Behavioral Sciences, Bilingual Review, Bilingual Research Journal, Bilingualism: Language and Cognition,* and *Southwest Journal of Linguistics.* He is the author of *A Minimalist Approach to Intrasentential Code Switching* (Garland, 1999), and editor of the forthcoming volume *Grammatical Theory and Bilingual Codeswitching* (MIT Press). He is chair of the organizing committee of the Fourth International Symposium on Bilingualism, to be held at Arizona State University in April, 2003. He may be reached at macswan@asu.edu.

Appendices

Appendix A
Weighted Mean Scaled Scores

Subject	Grade		1998		1999			2000		
			LEP	All	LEP	EP	All	LEP	EP	All
Reading										
	2	M	542.93	570.14	550.11	578.98	575.35	555.91	584.00	580.64
		SD	15.13	23.35	15.23	18.55	23.23	15.40	18.25	23.17

Grade		LEP	All	LEP	EP	All	LEP	EP	All
	N	2388	4843	2557	2557	3319	2741	2741	3356
3	M	567.19	599.53	573.53	606.94	604.06	578.52	611.32	608.33
	SD	13.51	25.97	13.00	19.15	25.17	12.50	18.37	24.57
	N	2445	4874	2608	2608	4936	2788	2788	4979
4	M	593.31	626.58	597.35	630.15	629.20	601.07	634.35	632.77
	SD	11.91	25.10	12.08	18.19	24.52	11.84	17.47	24.03
	N	2295	4837	2422	2422	4894	2606	2606	4958
5	M	610.03	643.35	612.81	646.04	644.92	615.14	647.72	646.84
	SD	9.96	22.42	10.05	15.66	21.79	10.00	15.05	21.36
	N	2147	4798	2293	2293	4868	2436	2436	4924
6	M	624.30	655.79	627.47	660.78	657.78	628.96	662.54	659.37
	SD	9.04	19.55	9.12	14.70	18.99	9.49	14.64	18.98
	N	1449	3395	1522	1522	3319	1646	1646	3356
7	M	633.17	670.85	635.71	679.00	672.17	638.05	680.54	674.06
	SD	10.34	20.31	9.75	15.49	19.61	10.85	15.72	19.74
	N	885	1776	941	941	1769	997	997	1797
8	M	649.52	684.52	651.42	692.03	686.13	653.21	693.55	687.63
	SD	9.09	18.32	8.37	13.81	17.75	9.04	14.19	17.79
	N	860	1803	919	919	1819	982	982	1851
9	M	650.43	683.79	651.81	690.86	684.25	652.97	691.93	685.67
	SD	7.81	116.91	7.52	13.96	16.83	7.31	14.10	16.73
	N	602	1288	641	641	1288	693	693	1321
10	M	654.50	689.40	656.25	697.05	689.85	656.93	697.56	691.05
	SD	8.65	16.74	7.81	13.84	16.50	7.88	14.29	16.70
	N	602	1440	614	614	1435	704	704	1479
11	M	662.13	697.18	663.57	703.39	696.76	664.35	703.50	697.81
	SD	9.35	16.35	8.26	13.26	15.83	8.29	13.91	16.07
	N	571	1395	598	598	1393	661	661	1446

Subject	Grade	1998		1999			2000		
		LEP	All	LEP	EP	All	LEP	EP	All
Math									
	2	M	547.65	564.58	555.79	572.58	571.91	562.12	579.37	579.14
		SD	15.75	20.96	15.97	19.20	21.07	16.47	19.05	21.38
		N	2540	4875	2555	2557	4914	2737	2741	4969
	3	M	571.11	590.43	579.49	598.43	598.36	587.14	606.81	606.51
		SD	15.58	22.36	15.30	19.45	21.95	15.25	19.07	21.66
		N	2493	4883	2602	2608	4938	2783	2788	4980
	4	M	592.14	614.08	597.92	618.28	618.95	603.46	625.57	625.71
		SD	13.03	21.56	13.42	18.09	21.24	13.43	17.77	21.23
		N	2385	4846	2419	2420	4902	2604	2606	4959
	5	M	614.86	638.55	619.22	641.50	642.36	623.74	646.64	647.75
		SD	11.84	21.09	11.99	16.85	20.73	12.15	16.63	20.92
		N	2209	4808	2286	2292	4871	2431	2434	4927
	6	M	629.16	656.14	633.79	661.83	660.34	637.24	665.71	664.57

110

Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
		SD	12.45	21.67	12.99	18.08	21.65	13.62	18.30	22.03
		N	1486	3400	1520	1522	3322	1644	1646	3361
	7	M	643.68	667.56	646.47	674.23	670.16	648.62	676.83	673.15
		SD	10.90	19.16	10.61	17.09	18.85	11.44	17.61	19.61
		N	886	1777	941	941	1772	996	997	1793
	8	M	652.85	676.41	655.18	682.99	679.36	656.96	685.46	681.71
		SD	11.70	18.74	10.83	16.65	18.80	11.43	17.09	19.10
		N	868	1803	912	917	1810	979	981	1847
	9	M	667.00	688.24	668.78	698.79	689.71	670.22	696.90	692.25
		SD	11.20	16.98	10.28	14.97	16.76	9.91	15.42	17.08
		N	617	1285	638	641	1295	690	693	1326
	10	M	677.47	694.70	679.93	701.79	696.78	680.26	702.34	698.09
		SD	12.17	15.66	11.17	14.53	15.76	10.81	15.02	15.98
		N	609	1438	609	614	1433	696	704	1478
	11	M	680.76	699.80	684.06	707.38	702.07	684.69	708.36	703.99
		SD	13.66	17.52	13.05	16.18	17.57	12.84	16.89	18.03
		N	575	1397	594	598	1392	658	661	1439
			1998		1999			2000		
Subject	Grade		LEP	All	LEP	EP	All	LEP	EP	All
Language										
	2	M	558.26	580.56	563.58	588.20	585.12	568.34	592.11	589.54
		SD	13.09	20.79	13.37	17.11	20.75	13.78	17.15	20.93
		N	2483	4868	2546	2556	4910	2730	2740	4969
	3	M	571.63	596.13	578.35	603.84	602.17	584.12	609.41	607.56
		SD	13.33	22.57	13.49	18.14	22.27	13.64	17.63	21.92
		N	2442	4873	2586	2601	4933	2771	2786	4977
	4	M	595.30	620.61	598.68	623.71	622.82	602.67	628.02	626.78
		SD	11.92	20.89	12.25	16.60	20.62	11.93	15.95	20.11
		N	2367	4839	2407	2413	4895	2596	2604	4960
	5	M	607.68	634.34	610.14	636.88	636.29	613.07	639.29	639.03
		SD	10.76	20.49	10.96	15.88	20.29	11.02	15.59	20.15
		N	2190	4805	2278	2291	4871	2424	2433	1927
	6	M	617.70	643.43	620.59	648.11	645.75	622.65	650.80	648.31
		SD	9.64	17.68	10.11	14.36	17.75	10.55	14.56	17.96
		N	1472	3394	1513	1521	3314	1634	1646	3354
	7	M	626.52	655.69	628.79	663.36	657.72	631.26	665.77	660.41
		SD	9.74	17.65	9.34	14.56	17.57	10.18	14.77	17.78
		N	883	1768	938	941	1768	991	996	1790
	8	M	632.29	661.89	633.81	669.37	664.12	635.63	671.49	666.25
		SD	9.07	17.89	8.74	14.86	18.05	9.28	15.21	18.22
		N	866	1799	913	918	1802	978	982	1843
	9	M	642.68	668.38	643.96	676.22	669.97	644.84	678.01	672.07
		SD	9.29	15.47	8.98	13.19	15.67	9.07	13.76	16.08
		N	608	1285	633	640	1280	688	693	1318

10	M	639.13	669.03	640.67	678.03	670.48	641.48	679.24	672.59
	SD	9.64	17.34	8.64	14.64	17.43	8.95	15.26	17.67
	N	597	1431	606	613	1425	693	703	1463
11	M	650.25	678.16	652.09	686.12	679.64	652.70	686.90	681.33
	SD	9.42	15.69	8.83	13.37	15.99	8.91	14.46	16.44
	N	565	1386	588	597	1383	658	661	1438

		1998		1999			2000		
Subject	Grade	LEP	All	LEP	EP	All	LEP	EP	All
Spelling									
2	M	533.17	558.19	542.89	568.99	565.13	550.35	575.46	572.04
	SD	19.86	23.42	19.94	19.41	23.26	20.75	19.27	23.49
	N	2502	4872	2553	2553	4912	2739	2739	4970
3	M	567.11	589.48	574.74	598.86	595.70	582.39	604.80	601.92
	SD	17.49	20.30	16.97	16.15	19.96	16.67	15.92	19.38
	N	2487	4883	2603	2603	4936	2785	2785	1981
4	M	583.53	612.64	588.39	617.46	615.69	593.67	623.30	620.95
	SD	14.27	23.21	14.58	17.86	22.60	14.21	17.27	22.14
	N	2384	4849	2420	2420	4900	2604	2604	4961
5	M	603.56	629.88	606.59	634.13	632.42	610.18	636.86	635.52
	SD	11.46	19.07	11.63	14.77	18.90	11.73	14.16	18.61
	N	2200	4810	2286	2286	4874	2430	2430	4929
6	M	611.54	642.94	615.15	649.85	645.85	618.38	653.33	649.06
	SD	11.76	20.18	12.39	16.11	20.07	12.68	16.17	20.33
	N	1486	3400	1521	1521	3323	1645	1645	3361
7	M	623.16	657.22	625.20	665.34	658.31	627.47	667.90	660.98
	SD	11.08	18.26	10.91	14.55	17.92	12.25	14.86	18.40
	N	887	1775	938	938	1771	995	995	1798
8	M	641.29	668.60	642.20	675.44	670.07	643.64	677.14	671.75
	SD	8.00	14.65	7.87	12.03	14.68	8.27	12.32	14.82
	N	870	1807	916	916	1816	978	978	1848

		1998		1999			2000		
Subject	Grade	LEP	All	LEP	EP	All	LEP	EP	All
Science									
9	M	651.24	670.30	652.79	675.05	671.21	653.62	675.96	672.39
	SD	6.60	12.30	6.05	10.35	12.00	5.95	10.47	11.98
	N	616	1281	639	640	1290	689	693	1325
10	M	656.39	676.94	657.96	682.40	678.00	657.81	682.15	678.20
	SD	7.20	13.22	6.04	11.54	13.07	6.03	11.85	13.15
	N	606	1428	609	614	1434	691	704	1471
11	M	659.77	681.99	661.60	687.48	683.23	662.08	687.60	683.98
	SD	7.41	13.66	6.79	12.15	12.00	6.47	12.50	13.80
	N	571	1388	593	598	1387	658	661	1436
Social Studies									

9	M	632.42	649.34	633.57	652.97	649.61	634.83	653.88	650.91
	SD	5.09	11.28	4.54	9.69	10.89	4.01	9.54	10.65
	N	614	1279	638	641	1283	689	693	1323
10	M	633.48	652.44	634.03	656.65	652.61	634.04	656.44	652.87
	SD	5.75	11.72	4.93	10.24	11.51	4.71	10.40	11.46
	N	604	1425	608	614	1428	693	704	1466
11	M	645.77	665.91	647.20	671.00	666.65	647.86	670.94	667.45
	SD	6.56	12.49	5.81	10.39	12.00	5.70	10.74	12.02
	N	573	1384	590	598	1379	657	661	1434

Appendix B
Weighted Mean Within-grade Gains

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Reading									
2		M	7.84	5.88	4.20	4.40	5.39	12.90	11.21
		SD	9.82	8.89	10.54	9.67	8.31	10.88	9.66
3		M	8.30	5.02	3.12	4.72	4.57	11.58	9.60
		SD	10.87	8.36	10.98	9.41	7.80	10.24	8.71
4		M	6.47	2.69	2.16	3.58	3.79	7.60	6.48
		SD	10.27	8.02	10.07	8.69	7.48	9.65	8.44
5		M	3.81	1.65	1.27	1.88	1.83	4.54	3.44
		SD	7.68	7.09	7.92	7.75	6.90	8.88	7.56
6		M	3.56	2.12	1.96	1.97	1.51	4.18	3.68
		SD	8.01	5.93	6.84	6.52	5.52	7.35	6.23
7		M	3.64	1.37	3.94	2.46	1.96	4.26	3.30
		SD	7.44	5.10	8.65	5.61	4.77	8.39	5.45
8		M	3.55	1.62	2.35	2.37	1.52	3.46	3.15
		SD	6.97	4.59	7.85	5.94	4.57	8.08	5.25
9		M	1.07	0.41	1.29	1.32	1.35	2.61	1.81
		SD	6.07	4.11	5.62	4.23	3.99	6.50	4.20
10		M	1.30	0.44	0.83	1.14	1.13	2.44	1.59
		SD	6.67	4.44	6.60	4.68	4.42	7.99	4.56
11		M	1.52	-0.36	0.91	0.98	0.92	2.31	0.60
		SD	7.28	5.15	7.38	5.30	5.26	7.65	5.23

Subject	Grade		1998-1999		1999-2000			1998-2000	
			LEP	ALL	LEP	EP	ALL	LEP	ALL
Math									
2		M	10.08	7.72	6.91	7.63	7.29	14.37	15.04
		SD	12.82	10.02	14.14	12.01	10.18	13.47	11.66
3		M	11.82	8.17	7.35	9.60	8.38	16.43	16.56

Grade		1998-1999 LEP	1998-1999 ALL	1999-2000 LEP	1999-2000 EP	1999-2000 ALL	1998-2000 LEP	1998-2000 ALL
	SD	13.23	9.87	12.92	10.59	9.38	13.08	11.03
4	M	7.84	4.88	4.95	7.33	6.91	11.30	11.83
	SD	10.33	8.71	11.39	9.51	8.35	11.08	9.69
5	M	4.90	3.78	4.49	5.68	5.34	8.44	9.09
	SD	9.10	8.31	10.03	8.84	8.18	10.37	9.39
6	M	4.66	4.30	3.67	4.89	4.11	7.55	8.49
	SD	10.06	7.54	9.66	8.61	7.51	9.51	8.51
7	M	3.80	2.60	2.04	3.13	3.08	4.61	5.69
	SD	7.73	5.25	7.72	6.30	5.37	7.01	5.97
8	M	4.09	2.97	1.92	2.73	2.40	4.19	5.38
	SD	7.16	5.04	7.88	6.85	5.27	7.03	5.93
9	M	1.71	1.46	1.68	2.53	2.52	3.52	3.97
	SD	5.75	4.30	4.96	4.60	4.23	6.05	4.54
10	M	2.36	2.14	0.68	1.24	1.25	3.16	3.40
	SD	5.82	4.01	5.28	4.48	3.99	6.50	4.33
11	M	3.42	2.35	0.88	1.92	1.83	4.30	4.22
	SD	6.64	4.68	6.30	5.23	4.82	6.69	5.14

Subject	Grade		1998-1999 LEP	1998-1999 ALL	1999-2000 LEP	1999-2000 EP	1999-2000 ALL	1998-2000 LEP	1998-2000 ALL
Language									
	2	M	5.69	5.85	3.82	4.07	4.55	9.95	9.46
		SD	9.25	8.24	11.01	9.05	8.07	10.34	9.26
	3	M	8.34	6.42	3.54	6.02	5.62	12.88	12.00
		SD	10.92	8.32	10.58	9.77	7.96	10.59	9.02
	4	M	5.56	2.24	1.83	3.43	4.15	7.30	6.35
		SD	9.62	7.56	10.09	8.41	7.43	9.93	8.20
	5	M	3.25	1.99	2.02	2.13	2.66	4.85	4.59
		SD	8.57	7.66	8.71	8.25	7.56	9.74	8.17
	6	M	3.08	2.37	1.99	3.19	2.49	4.50	4.93
		SD	8.78	6.62	7.80	6.77	6.31	7.92	6.96
	7	M	3.48	2.04	3.11	2.75	2.76	4.24	4.80
		SD	7.50	5.15	8.82	5.84	5.02	7.66	5.58
	8	M	2.92	2.24	2.56	2.49	2.18	3.31	4.41
		SD	5.97	4.96	8.55	7.37	5.10	7.55	5.63
	9	M	1.01	1.54	1.11	2.23	2.10	2.41	3.67
		SD	5.85	4.21	5.43	4.54	4.24	6.72	4.52
	10	M	1.24	1.46	1.24	2.06	2.05	2.53	3.53
		SD	6.08	4.89	6.08	5.36	4.95	7.38	5.06
	11	M	2.00	1.56	0.61	1.68	1.58	2.62	3.21
		SD	6.30	4.67	6.59	5.22	5.03	7.01	5.09

Subject	Grade		1998-1999 LEP	1998-1999 ALL	1999-2000 LEP	1999-2000 EP	1999-2000 ALL	1998-2000 LEP	1998-2000 ALL
Spelling									
	2	M	9.49	7.46	7.70	5.89	6.96	16.91	14.36
		SD	12.38	9.80	11.84	25.10	9.46	13.83	10.96
	3	M	7.80	6.61	8.06	5.55	6.41	15.76	13.01
		SD	10.99	8.59	10.66	20.02	8.50	12.16	9.46
	4	M	4.44	3.12	5.68	4.91	5.48	10.14	8.60
		SD	9.68	8.45	9.91	20.82	8.17	10.43	9.14
	5	M	2.80	2.59	3.25	2.12	2.99	6.12	5.56
		SD	9.04	7.36	8.60	21.36	7.21	9.80	7.91
	6	M	3.05	2.98	2.94	2.44	3.19	6.14	6.26
		SD	8.75	7.17	8.71	24.24	6.99	9.50	7.75
	7	M	1.65	1.06	1.98	2.42	2.78	3.68	3.86

114

	SD	7.25	4.89	7.67	14.15	4.92	8.21	5.42
8	M	0.83	1.44	1.23	0.97	1.70	2.20	3.14
	SD	5.91	4.49	5.99	22.11	4.57	6.99	5.25
Science								
9	M	1.38	0.87	1.06	0.89	1.15	2.56	2.03
	SD	5.21	3.60	4.59	12.65	3.28	5.53	3.69
10	M	1.38	1.10	0.01	-0.35	0.14	1.55	1.24
	SD	5.34	4.03	4.60	17.65	3.76	5.99	4.14
11	M	1.86	1.33	0.46	0.20	0.65	2.34	2.02
	SD	5.25	4.42	5.10	16.08	4.39	5.65	4.56
Social Studies								
9	M	1.08	0.22	1.46	0.93	1.26	2.58	1.49
	SD	4.53	3.10	3.77	12.45	2.97	4.36	3.20
10	M	0.35	0.14	0.09	-0.19	0.23	0.61	0.37
	SD	4.56	3.43	4.05	15.78	3.22	5.09	3.43
11	M	1.58	0.77	0.78	0.39	0.71	2.16	1.52
	SD	5.46	4.21	5.30	14.81	3.95	5.57	4.26

Appendix C
Weighted Mean Cohort Gains

Subject	Grade		1998-1999		1999-2000			Grades	1998-2000	
			LEP	ALL	LEP	EP	ALL		LEP	ALL
Reading										
	2 to 3	M	30.59	34.00	28.7	34.21	32.66	2 to 4	58.33	62.19
		SD	9.51	8.48	9.83	9.36	8.61		11.01	9.20
	3 to 4	M	30.27	29.54	27.87	27.76	28.35	3 to 5	47.99	47.05
		SD	8.89	7.28	8.97	7.85	7.11		10.06	8.93
	4 to 5	M	19.51	18.31	17.91	16.75	17.53	4 to 6	38.75	33.53
		SD	8.43	7.06	8.35	7.49	6.79		10.94	9.77
	5 to 6	M	19.75	15.72	18.88	14.91	15.92			
		SD	7.41	7.24	8.23	7.27	7.00			
Math										
	2 to 3	M	31.61	33.74	31.76	35.30	34.37	2 to 4	55.96	60.70
		SD	12.23	11.03	12.61	12.02	11.20		13.21	11.64
	3 to 4	M	26.81	28.30	24.38	27.49	27.07	3 to 5	52.83	57.00
		SD	11.02	9.65	11.12	10.74	9.76		12.85	11.17
	4 to 5	M	27.13	28.26	26.07	28.91	28.73	4 to 6	50.46	53.54
		SD	9.43	8.16	9.53	9.03	8.08		12.65	10.95
	5 to 6	M	22.79	24.58	22.60	25.67	25.43			
		SD	9.36	8.96	9.64	9.32	8.80			
Spelling										
	2 to 3	M	41.78	37.55	40.24	35.15	36.66	2 to 4	61.12	62.30
		SD	11.00	9.19	10.68	9.88	9.17		12.79	9.55
	3 to 4	M	21.57	26.03	19.51	26.15	24.82	3 to 5	43.41	45.79
		SD	10.08	8.04	10.13	8.91	8.21		11.73	8.38
	4 to 5	M	23.27	19.82	22.04	18.34	19.77	4 to 6	37.95	37.41
		SD	8.70	7.91	9.09	8.46	7.72		11.58	9.49
	5 to 6	M	13.88	17.32	13.88	18.61	18.04			
		SD	8.89	7.47	9.28	7.96	7.54			
Language										
	2 to 3	M	20.06	21.70	21.02	22.48	22.29	2 to 4	44.52	45.78
		SD	9.40	8.53	9.64	9.64	8.52		10.63	9.00
	3 to 4	M	27.13	26.50	24.71	23.62	24.30	3 to 5	41.50	42.56
		SD	9.21	7.92	9.35	8.61	7.77		10.98	9.03
	4 to 5	M	14.83	15.69	14.54	15.84	16.14	4 to 6	31.03	29.07

			8.73	6.82	8.62	7.56	6.74		10.67	8.85
5 to 6	M		15.37	12.84	15.01	12.89	13.72			
	SD		8.08	7.70	8.56	8.11	7.58			

Appendix D
Unweighted Mean Scaled Scores

Subject	Grade		1998		1999			2000		
			LEP	All	LEP	EP	All	LEP	EP	All
Reading										
	2	M	547.69	572.91	554.73	570.92	578.67	560.34	578.08	583.94
		SD	17.95	23.06	17.70	19.18	22.82	17.47	19.08	22.65
		N	2388	4843	2557	2557	3319	2741	2741	3356
	3	M	571.02	602.21	577.13	596.33	607.21	581.95	605.26	611.58
		SD	15.33	25.72	15.24	19.47	24.87	14.44	18.95	24.2
		N	2445	4874	2608	2608	4936	2788	2788	4979
	4	M	596.17	628.96	600.18	616.57	631.78	604.38	624.12	635.56
		SD	13.98	24.82	13.71	18.02	24.31	13.74	17.46	23.72
		N	2295	4837	2422	2422	4894	2606	2606	4958
Math										
	2	M	550.59	566.41	558.44	576.57	573.95	565.5	582.29	581.36
		SD	18.48	21.18	18.75	18.52	21.26	18.98	18.1	21.31
		N	2540	4875	2555	2557	4914	2737	2741	4969
	3	M	574.02	591.96	582.39	604.47	600.25	590.49	609.41	608.48
		SD	17.78	22.41	17.83	19.09	22.09	17.56	18.15	21.48
		N	2493	4883	2602	2608	4938	2783	2788	4980
	4	M	594.48	615.47	600.48	627.87	620.52	606.75	632.41	627.5
		SD	15.32	21.45	15.29	17.9	21.18	15.43	17.15	20.98
		N	2385	4846	2419	2420	4902	2604	2606	4959
Language										
	2	M	561.73	582.3	566.96	586.13	587.17	571.59	590.57	591.59
		SD	15.57	20.55	15.57	17.03	20.31	15.7	17.1	20.41
		N	2483	4868	2546	2556	4910	2730	2740	4969
	3	M	574.98	597.75	581.86	601.61	604.16	587.39	607.58	609.49
		SD	.15.5	22.33	15.78	18.01	21.93	15.71	17.47	21.53
		N	2442	4873	2586	2601	4933	2771	2786	4977
	4	M	597.9	622	601.72	621.92	624.21	605.9	626.58	628.25
		SD	13.9	20.48	13.97	16.46	20.22	13.68	15.68	19.61
		N	2367	4839	2407	2413	4895	2596	2604	4960

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

Education Policy Analysis Archives

Reactions to Bolon's "Significance of Test-based Ratings for Metropolitan Boston Schools"

Stephan Michelson
Longbranch Research Associates

Abstract Several concerns are raised abut the procedures used and conclusions drawn in Craig Bolon's article "Significance of Test-based Ratings for Metropolitan Boston Schools" published in this journal as Number 42 of Volume 9.

Craig Bolon introduces his article about the Massachusetts Comprehensive Assessment System (MCAS), tenth grade mathematics tests, with a non-sequitur. "The state is treating scores and ratings as though they were precise educational measures of high significance," he tells us, but "statistically they are not." Nothing Bolon does leads to this conclusion. We do not know how "precise" these measures are, by any criterion. We do not know how "significant" they are, by any criterion. One wants the conclusions of an article to represent its contents. Bolon's do not. Whether scores measure something consistently is called "reliability," and Bolon's exposition implies highly reliable tests. [1]

Whether scores measure something else, something besides themselves, is called "validity." Bolon has not said what concept success on these tests is supposed to imply, or that they don't. He has not determined that these scores are invalid against some criterion. He finds them correlated with town income, a not surprising result that we have lived with for 35 years, since the publication of The Coleman Report. [2] Nothing in that correlation informs us about the validity of these test scores, their ability to predict some other attribute of the students who take them.

Bolon concludes that student characteristics, such as limited English proficiency, "all failed to associate substantial additional variance" over and above that accounted for by town income. As I will explain, that criterion is not helpful in determining if a variable should be in a model explaining test scores. And at any rate, it is incorrect. Bolon reaches that conclusion by dropping Boston schools from his data, weighting observations by size of school, and failing to take advantage of the very information he provides. He also claims that the data cannot display "performance trends," though he has not established that there are trends that the data fail to capture. The incentive systems the state will associate with school performance are not yet in place. Bolon cannot therefore say that schools have failed to react to them.

Finally, Bolon tells us that per pupil school expenditure *is* not well related to school performance. Or perhaps he is saying that is true over and above town income. It appears to be the lack of correlation between test scores and school expenditure that leads Bolon to conclude that schools cannot raise themselves when found to be at a low level.[3] In fact, per pupil expenditure is highly correlated with test score, and with income, also.

This comment is only about statistical analysis, about deriving information from data. I say "information" rather than "fact" because all statistical results are probabilistic, associations one can choose to accept or not on many grounds. On the surface that

is what Bolon offers, statistically derived information. But his conclusions do not follow from his data. His procedures are suspect and his interpretations are incorrect. Surely the discussion requires better statistical analysis.

Data

Bolon has put together data from various sources, describing 47 schools in 32 towns around Boston. For this he deserves much credit, although one should not expect much from these data. Per pupil expenditure, for example, is presumably over an entire school system, and includes many items that no one would think to associate with a mathematics test score. Only if all students taking the test have always been in the school system in which we see them, at the tenth grade--and assuming that last year's expenditures are representative of the last ten years of expenditures--would system-wide expenditure be a reasonable measure of resources expended on these students.

Within his data, Bolon recognizes that three Boston schools, entrance to which is predicated on an exam, "draw away many Boston students who tend to score well on achievement tests...." [4] But he neither includes that fact in his statistical study, nor does he investigate the effect of this "creaming" on other Boston schools. Let us do so quickly.

Boston schools, on average, do not score differently from non-Boston schools. Non-Boston schools do score higher (226.6 to 212.5 in 1999), but a difference this large or larger would be expected to occur by chance more than forty-four times in one hundred in a world in which the actual difference were zero. In Boston, the three exam schools out-score the ten non-exam schools (238 to 204.9) a difference that we would expect to occur by chance, assuming they were equal, fewer than two times in one hundred. [5] So a statistical examination of test scores should account for both those Boston schools that contain the cream (the variable I call "exam"), and those that contain only the remaining skimmed milk (the variable "bosnoex").

Bolon includes schools that "provide vocational education in the same facilities as academic programs." [6] He notes that "most students in vocational programs receive lower MCAS scores than students in academic programs" [7] when they are in separate schools, but nonetheless takes no account of vocational students when mixed with academic students. Bolon indicates this mix with an asterisk attached to twelve schools, none of them in Boston. I created the variable "voc," with a value 1 for schools that have a vocational program, 0 for schools that do not. A better variable would be the percent of students in each school who were "vocational." An even better variable would be the percentage of *test takers* in each school who could be categorized that way. If the indicator variable has a negative effect and is statistically unlikely from chance, when we assume it is structurally unrelated, then perhaps it is telling us something about these data that no other variable captures. It would not be *new* information, but how would one justify not taking into account old, correct information in asking what new information these data contain?

Replication

Furthermore, there is the problem of "specification error." Although this seems like a technical term, it is part of a larger inquiry about the language we use to describe our findings. The "specification" of a single equation regression is simply the variables, and their forms (linear, logarithmic, quadratic, etc.), including weights. Some variables might be important to specification even though their associated probability is high, because their inclusion changes the coefficients of other variables. One variable, though Bolon does not use it, does show some specification effect.

Bolon implies that school size is such a variable, but rather than including it, he weights observations by it. He is wrong on both counts: School size is not only not important for specification, weighting by it is incorrect. Weighting implies that there is more information in a large school than a small school. [8] If size affects score, we should determine that relationship directly (as a variable in the equation). Otherwise, Bolon's observations are of schools, each one telling us as much as any other.

To this point it appears that Bolon has not taken advantage of his own data, failing to define three variables (exam, bosnoex and voc) that his own discussion implies should be important. If I am going to explain the consequences of these omissions, I should first replicate his results. Only if I do so can I assert that differences between us are due solely to the changes in specification that I will offer. It turns out that there is more amiss here than just the variables.

I placed my coefficients and standard errors into Excel, and let it round them to three places. Bolon uses what to me is awkward language for R^2, the most common statistic emanating from a regression equation. He says that a model "associates X percent of the statistical variance." Perhaps he is avoiding the more usual term, "explains." I am sympathetic to that desire, but "associates" is usually followed by the word "with." I find his phrasing unintelligible. If one were asked to estimate the average school test score of each school, knowing nothing about schools, the best strategy is to estimate the mean of all schools every time. R^2 essentially tells you how much better than that you can do (how much closer to real values your estimates are) if you generate an estimated score from the regression equation. R^2 can vary from zero to 1.00.

In an exercise in which Bolon will use different numbers of explanatory variables, he should report the adjusted R^2, which makes the statistic pay a penalty for using more variables. The easy strategy to increase R^2 is to use all the variables you have. R^2 itself can never decrease when variables are added, an important item to remember as Bolon changes data on us. Adjusted R^2 can decrease, telling you that you have paid too high a price for whatever information your last variable has added. It can even become negative. I will return, below, to how Bolon appears to be using R^2 to create his specification. In Table 1, a replication of his Table 2-6, to the three places Bolon reports, we agree exactly.

120

Table 1

Replication of Bolon Table 2-6

	Coefficient	Standard Error
constant	229.4	1.936
perafro	0.047	0.104
perasian	0.347	0.154
perhisp	-0.002	0.183
perlimit	-0.637	0.217
perlunch	-0.174	0.157
R^2	0.6697	adjusted 0.6294

The prefix "per" represents "percent." I believe we are looking at the percent black, Asian and Hispanic in the school, not necessarily taking the exam. Similarly, "perlimit" is percent of the school with limited English fluency (whether defined consistently from school to school I cannot say), and "perlunch" is percent receiving free or reduced cost lunch. A variable we will encounter below, "percapy," is not a percentage. It is per capita income in 1989 (1990 Census data). Short of administering a questionnaire to each student--and then dealing with the accuracy of the responses--these are the kinds of variables we have in education research, somewhat distant from what we would like to measure.[9]

I will present R^2 and adjusted R^2 from left to right in the following tables, without including the word "adjusted." I could show a similar table (replicating Bolon's) for his Table 2-7. In Table 2-9, Bolon and I disagree in the last decimal place shown of four coefficients. It is possible that Bolon has data to more decimal places than he has reproduced. [10] This possibility is suggested by coefficients for per capita income in his equation in Table 2-10. Mine is 1.108, his is 1.104. Not consequential, but not a rounding difference, either.

A Change of Method

I think we can agree that my results are what Bolon would derive from the data he has provided. Just after his Table 2-13, Bolon makes a turn that he does not justify, and that I see no reason for. He tells us in Table 2-13 that the unadjusted R^2 for the three factor equation in Table 2-10, in 1999, is .80. That is, the residual variance of the difference between school scores and his estimate thereof is 20 percent of the value it was when he estimated every school to have the mean score of all schools. He has, in more traditional language, "explained" 80 percent of the variance in scores. Yet he will tell us, in Table 2-15, that a two factor model produces an R^2 of .86, and that he needs only per capita income to generate an R^2 of .84. Higher unadjusted R^2 from fewer variables? Something else has changed.

Table 2

All Schools Replication of Bolon Table 2-14

	Michelson		Bolon
	Coefficient	Standard Error	
constant	209.9	4.717	
perlimit	-0.614	0.091	
percapy	0.999	0.219	
R^2	0.7409	0.7291	

weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	211.5	4.863	201.5	2.934
perlimit	-0.629	0.108	-0.325	0.136
percapy	0.949	0.224	1.307	0.126
R^2	0.6802	0.6657	0.86	

One difference lies in the statement just before Table 2- 14, that from now on regressions are calculated "with schools weighted by numbers of test participants." Bolon does not tell us why he has changed his specification that way. If this is how regressions should be run on these data, why did he not start out doing so? Nor does this change in procedure alone allow me to replicate his results.

Table 3

All Schools Replication of Bolon Table 2-16

	Michelson		Bolon	
	Coefficient	Standard Error		
constant	190.4	5.253		
percapy	1.715	0.270		
R²	0.4734	0.4617		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	195.0	5.229	197.0	2.395
percapy	1.539	0.263	1.465	0.114
R²	0.4316	0.4189	0.84	

In Tables 2 and 3 I have weighted by the adjusted number of test takers, as listed in Bolon's Appendix Table A3-1. Weighting makes little difference in the coefficients or the R^2. Weighting surely should be justified, which I believe it cannot be in this context, but neither is it of any consequence.

As implied by my table titles, and R^2 far below the values Bolon reports, what Bolon has done is drop some schools. Boston schools. He does so, partly, "because the students who score well on school-based standard tests are selected for admission to the three exam schools," and for other reasons mostly confined to a footnote. [11] As we know, the exam schools can easily be represented by a variable. He notes that attributing the same per capita income to all ten schools in Boston does not allow that variable to explain differences in test scores among those schools. True, but deleting variation because you cannot explain it is guaranteed to both raise R^2 and leave you ignorant. For Bolon to conclude, later, that per capita income is the only factor needed to explain most variation in test scores, is to ignore that he has deleted 21 percent of his data to achieve that result.[12] It is not true of his initial data set. He does not feel compelled to make the same adjustment in Lynn, Newton or Quincy, all of which have the same per capita income applied to more than one school. Nor is it necessary to adjust the data at all.

One adjustment Bolon might have made is to have one observation per town, by creating a weighted average of scores (as he tells us how many test takers there are per school). In fact, I think there is no school effect to be found in these data—though there are pupil effects—and therefore a per town analysis would have been best. I will return to that view. Another answer is to use the variables I suggested above, which assume that exam schools will have higher than average scores because they have selected students on that basis, and that Boston's non-exam schools will have lower than average scores because the students they would ordinarily enroll have been snatched from them by the exam schools. First, I will show that dropping the Boston schools, not weighting, is the key to Bolon's analysis:

Table 4

Non-Boston Schools Replication of Bolon Table 2-14

	Michelson		Bolon	
	Coefficient	Standard Error		
constant	202.7	3.258		
perlimit	-0.354	0.147		
percapy	1.264	0.141		
R²	0.8364	0.8259		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	201.5	2.940	201.5	2.934
perlimit	-0.325	0.136	-0.325	0.136
percapy	1.309	0.126	1.307	0.126
R²	0.8619	0.8530	0.86	

Table 5

Non-Boston Schools Replication of Bolon Table 2-16

	Michelson		Bolon	
	Coefficient	**Standard Error**		
constant	197.5	2.613		
percapy	1.451	0.126		
R²	0.8060	0.7999		

weighted	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**
constant	196.9	2.399	197.0	2.395
percapy	1.467	0.115	1.465	0.114
R²	0.8364	0.8313	0.84	

As earlier, here in Tables 4 and 5, comparing equations vertically on the left, there is little difference between weighted and non-weighted results, again raising the question why Bolon bothered to change his method at this point. Reading the weighted equations left to right on the bottom, there are some differences between us in the last decimal place that cannot be attributed to output rounding. Essentially I replicate Bolon, which means that he has decided that Boston schools do not contribute to our knowledge about test scores in Boston-area schools.

He is quite wrong about that. Consider this five factor "model," which I present unweighted:

Table 6

Michelson's 5-Factor Model

	Coefficient	Standard Error	p>|t|
constant	205.3	3.461	0.000
perlimit	-0.289	0.110	0.012
percapy	1.191	0.149	0.000
exam	16.238	2.670	0.000
bosnoex	-10.307	3.111	0.002
voc	-3.981	1.564	0.015
R²	0.9051	0.8935	

I have added a column. I do not find the standard error a very informative statistic. Without the t-distribution at hand, one cannot translate the coefficient and standard error to probability. Many people report the t statistic, which is the ratio of the coefficient to its standard error, but it suffers from the same problem. I prefer to report the probability itself, as that is what everyone receiving other information is trying to derive.

Specification

Bolon is much too attached to reporting "significance, and eliminating variables "not significant at p < .05." That is what I call a "mechanical" approach to statistical inference, one a machine can do as well as a human. Probability is not like an on-off light switch. It is more like a dimmer. There is more or less of it. p = .06 should not be set aside as "not significant." It requires interpretation. I would accept a probability for the vocational variable higher than for other variables, because it masks the variation (the percent of exam takers who are vocational) that it should have. The equation in Table 6 does not need this explanation. [13] It would have been a superior model with higher probabilities, but its probabilities meet Bolon's apparent criterion.

The three classification variables included here, derived from Bolon's data but not presented as such by him, are extremely powerful explainers of variation in school test scores. My equation achieves a higher R^2, even adjusted, than any unadjusted R^2 Bolon reports. I don't say that should be the single criterion for choosing a model, but it does seem to be one of Bolon's, and is a good starting point. Retaining all the data and achieving a higher R^2 than Bolon's model did, after his dropping data for the sole purpose of increasing R^2, surely argues that my model is more informative than his.

If statistical comparisons are too ornery--I often think they are--a picture should make my point. Let's compare the ability of Bolon's "one-factor model" and my "five factor model" to predict the data, to estimate the actual data points from applying the equation to input data.

In Figure 1, test scores from 1999 are plotted for all 47 schools, against per capita income. Triangles represent data points. Estimated points are implied by the lines that join them. The solid line is the one Bolon would have us believe best describes the

data, a regression on per capita income alone. The dashed line follows estimates made from my five factor equation, Table 6. Does Bolon's one factor model capture the essence of this data set? I don't think so.

Figure 1
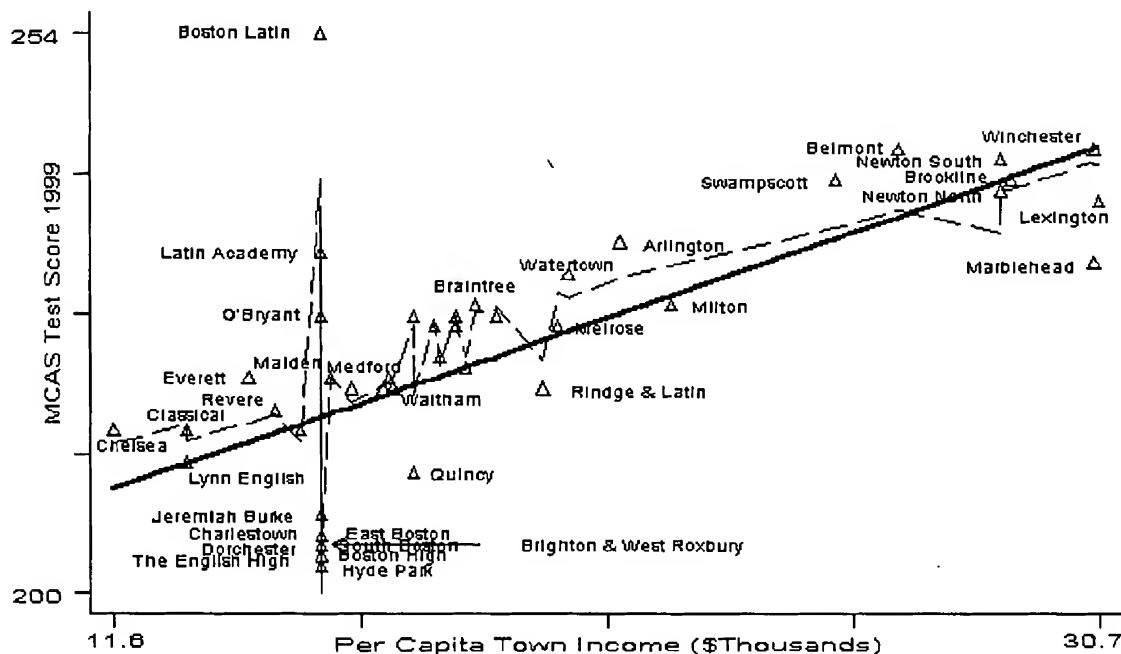
Boston schools lie on top of each other, because per capita income is defined only by town. Bolon's regression line, of course, cannot distinguish among them. However, my regression distinguishes all except Boston Latin very well. It is true that the non-exam Boston schools cannot vary in per capita income, but neither do they vary much in test score. Although not always in the same place, the ten Boston non-exam schools had the ten lowest school averages in all three years of Bolon's data. Boston Latin always had the highest score. Bolon's equation—or, to be more precise, my equation using Bolon's specification on all schools— under-estimates scores of low income towns and over-estimates scores of high income towns. In short, he presents not only an inaccurate but a biased view of the net relationship between per capita income and test score.

Besides looking to coefficient probability and R^2, how does Bolon (apparently) select his model? In Table 2-12 he presents R^2 for each combination of three factors. In Table 2-15 he asks how much increase in R^2 he can obtain from adding alternative other variables to his two factor model (now based on only 34 schools, weighted). He tells us that one factor, per capita income, gives him an R^2 of .84, so why bother with limited English proficiency when it increase R^2 to only .86? This is a political decision. Bolon wants to say that a model with no variable describing school or child characteristics tells us as much as we can know about these data. He wants to argue that school test data that relate only to town income cannot be valid, though this is not how validity is measured. There is no statistical justification for dropping other variables.

Table 7

Increments to R² in
Michelson's Five Factor Model

Model	R²	Adjusted R²
Michelson's 5-factor	0.9051	0.8935
drop percapy	0.7575	0.7344
drop exam	0.8195	0.8023
drop Bosnoex	0.8797	0.8682
drop perlimit	0.8891	0.8785
drop voc	0.8901	0.8797
Just percapy	0.4734	0.4691
Just exam	0.1012	0.0812
Just Bosnoex	0.5402	0.5300
Just perlimit	0.6181	0.6097
Just voc	0.0061	-0.0160

In Table 7, I show, first, how much the addition of each variable in my Five Factor Model adds to R^2. The number after "drop" is the R^2 from the remaining four variables. Thus, for example, per capita income adds .1476 to R^2, because R^2 is .7575 without it, and .9051 with it. Per capita income alone generates an R^2 of .4734, from the bottom half of Table 7. In other words, .3258 of the variation "explained" by per capita income is also explained by other variables. That is, 69 percent of the apparent explanatory power of per capita income may or may not be due to per capita income. It is "shared" with other variables. One might say that the variable "exam" explains little--an R^2 = .10 when it is the only variable in the equation--but only 15 percent of that is shared with other explanatory factors. Limited English proficiency explains more than any other single variable, but shares 97 percent of that variation with other variables. Is this a reason to drop this variable?

Not statistically. When Bolon drops perlimit, he allocates its shared explanatory variation to the remaining variables, in his case to percapy, by fiat. In the regression context, we just do not know whether it is per capita income or English proficiency, or both, or neither (they may both be proxies for something else) that is associated with variation in test scores. We let the multiple regression statistics tell us whether adding perlimit is worth the price of using up a degree of freedom. The variable describing limited English proficiency belongs in the model, even though 97 percent of the variation in test score that it explains can be explained by other variables. We can believe, with about as much confidence as one ever can get from data such as these, that where there are more students with limited English proficiency, the school's average test score will be lower.

The only exception to this arithmetic, where adding the R^2 of the equation with just one variable to the R^2 of the equation with all other variables produces a larger R^2 than we know all five produce, is vocational education. Adding its R^2 alone to the R^2 without it falls short of the R^2 we know we get from five factors. Vocational education displays a specification effect. Its presence in the equation increases the contribution to R^2 made by other variables. Bolon should have defined and utilized a "voc" variable regardless of its contribution to R^2.

Interpretation

We might infer that it is the students with limited English proficiency themselves who are causing a decline in the average, through their low scores. That is an inference Bolon is careful not to make. There is such a thing as the "ecological fallacy," or the "Simpson paradox," in which group data lead to incorrect inferences about individuals. But fear of that mistake should not deter someone from making an informed judgment. Studies in the 1960s and 1970s showed that states with more blacks had lower income, and states with higher education levels had higher income. It was not wrong to infer that blacks earned less than whites, and that higher educated people earned more than lower educated people, even though it *was* wrong to infer (at that time, though many did) that if black education levels rose they would earn considerably higher incomes. I think Bolon's data is quite sufficient to ask if the MCAS test is fair to persons with limited ability in English (and, presumably, Spanish, as tests are available in that language).

The same is true with vocational students. There could be many reasons why schools with vocational programs score lower than strictly academic schools, without the vocational students being the direct cause. But, armed with his other study of strictly vocational schools, Bolon surely can infer that it is the vocational students themselves who are scoring lower. This raises the question: What purpose is served by vocational students taking this academic test? Bolon's study does *not* answer this question, because it is not a validity study. But this is an example of the kind of useful inquiry Bolon could have provided.

We can conclude that two student descriptive variables appear important, an interesting if expected finding from aggregate data of this sort: limited English language ability and a vocational curriculum. In addition, we can conclude that creaming does matter, in the sense that concentrating all the best students in three schools concentrates the worst students in the remaining schools available to them. Ultimately, we can interpret these data as telling us nothing about schools, but much about students.

It was conventional wisdom in 1955, when I graduated from Brookline High, that aside from Boston Latin, one could find the

best greater Boston area students in Brookline and Newton, and maybe Lexington, largely because their parents had moved to these towns for that reason. Ten years later research declared what we knew to be true, though it was "controversial" in professional ranks: high scoring schools were so because they had high scoring students. And high scoring students had two characteristics: By and large they came from wealthier households, and they associated with each other in wealthy towns. Bolon's finding to this effect is surely correct, and welcome, but not new information, and hardly impugning the MCAS test on which it is based.

Residuals

Bolon estimates his one-factor model in three years, and concludes "that schools scoring higher than predicted tend to increase scores in successive years, and schools scoring lower than predicted tend to decrease scores." [14] This is not a correct interpretation from his own data and results. Residuals do not show the highest scoring schools scoring even higher from year to year, or the lowest scoring schools scoring even lower. The correlation of residuals means that they are in the same place every year. Furthermore, 1998 scores do not predict 1999 scores for the lowest scoring schools. The lower scoring schools have randomly higher or lower scores the next year. Although 1998 scores appear to predict higher 1999 scores for the highest scoring schools, that result is due to Boston Latin only. Except for Boston Latin, the higher scoring schools in 1998 have somewhat lower scores in 1999.

Figure 2 is a plot of residuals in 1999 against residuals in 1998, from my 5-factor model:

Figure 2

Five Factor Model Residuals, 1998
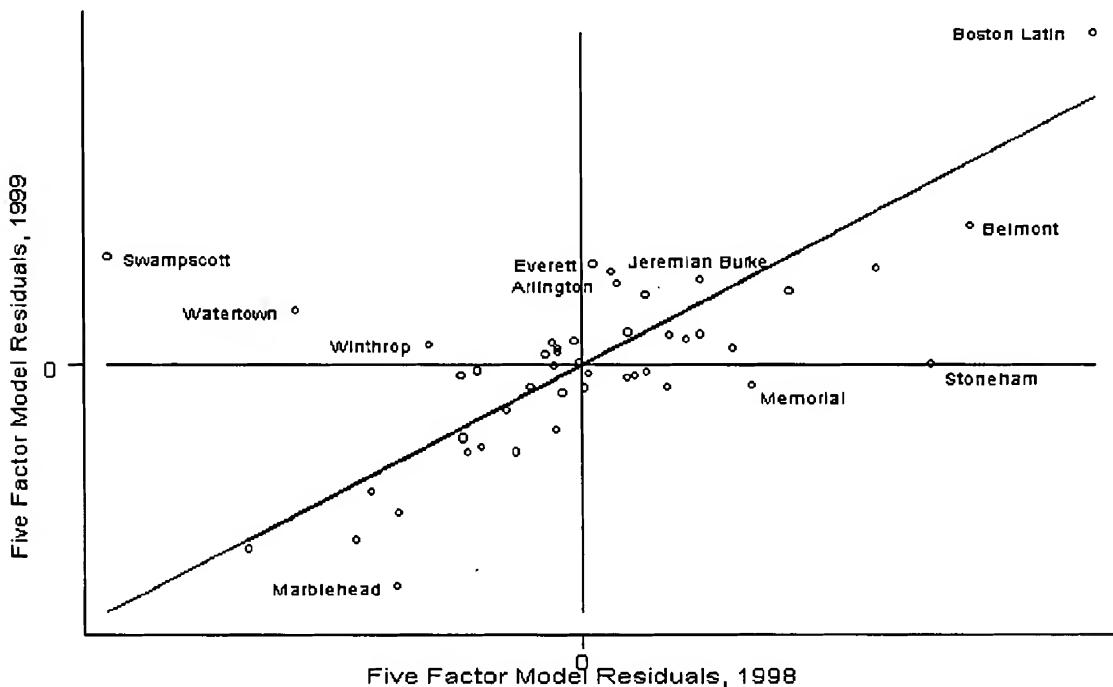
As I did also in Figure 1, I have named some schools, the principle being to inform but avoid clutter. In general, schools line up the same way every year. The residuals from 1998, for the top fifteen schools, predict the residuals from those schools in 1999 with a very small constant and a coefficient of .995. The residuals from the bottom fifteen schools in 1998 predict their 1999 residuals with a zero constant and a coefficient of 1.055. That is, we can predict the residual for the top and bottom schools almost precisely from the prior year data. The same does not hold true in the middle. Overall the residuals are correlated, as we would expect them to be:

Table 8

	resid98	resid99	resid00
resid98	1.0000		
resid99	0.6063	1.0000	
resid00	0.6241	0.6836	1.0000

More importantly, raw scores are also correlated across time, as indicated above in noting that the Boston non-exam schools always score at the bottom. Figure 3 is a picture of the ranks of the scores, from lowest to highest, in 1999 and 1998:

Figure 3
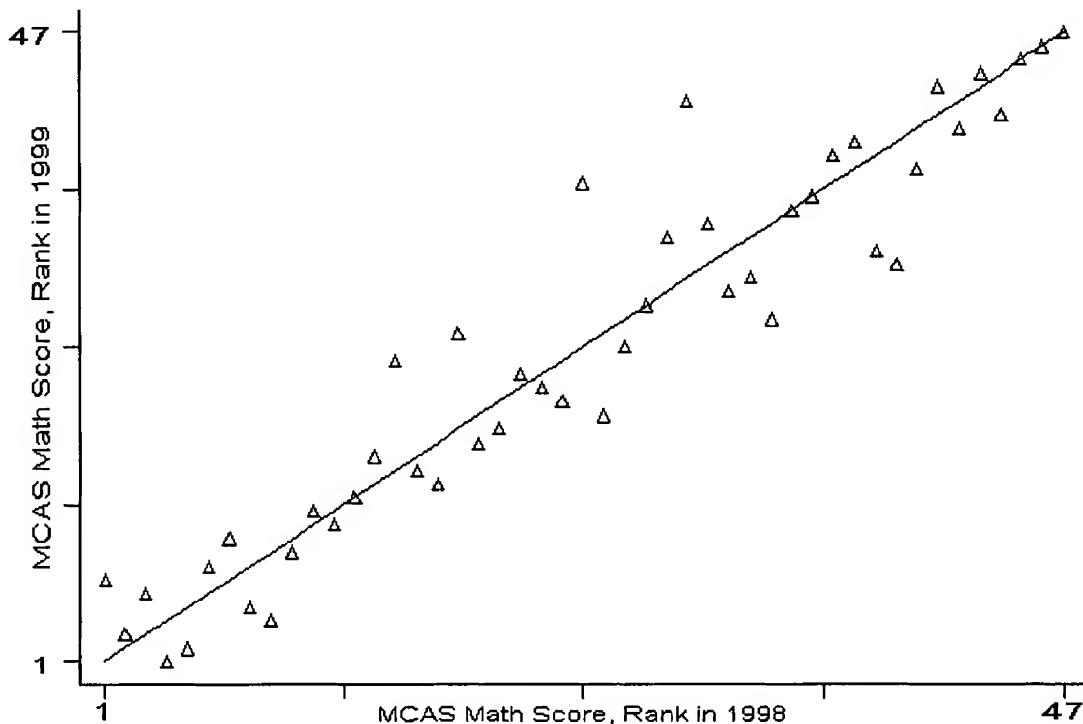
The best schools--which is to say the best students--are consistently best, and the worst are consistently worst, and in the middle there is some moving around, but the middle schools seldom appear best or worst. This is startlingly true, and largely unrecognized, of most reliability measures. A test with a high reliability can be very poor in the middle, where it is used to hire or not, to pass or to fail. It is likely that this same variability applies to individual students, and that reliability figures offered in support of the MCAS test do not describe the unreliability at exactly the pass-fail scores where it is most important to be correct.

The Town View

As noted above, another solution to Bolon's dilemma, that some variables described towns, not schools, would have been to aggregate all data by town. The assumption that would allow one to do this is that there is no school effect, for example that Boston Latin students score highest because they were pre-selected, not because they were better educated. Nothing in these data argues against that view.

Bolon chose a different route:[15]

> My summary analysis is based on a one-factor model for metropolitan Boston communities that each operate only a single academic high school, weighted by numbers of students tested. The effects study showed that "Per-capita community income (1989)" was the dominant factor in predicting 1999 school-averaged tenth-grade MCAS mathematics test scores. All other factors made only small contributions to predictions with much lower significance.

That is, Bolon's final comments are based on 60 percent of his initial data, 28 observations. It is hard to take Bolon's conclusions as seriously generalizing the results of a study of 47 schools from 32 towns. I will suggest below that only one town--not one selected by Bolon--might have been dropped to increase our understanding of existing relationships.

One of the interesting results one can obtain from analyzing towns is that property value (measured per capita in Bolon's data) appears to be a function of both income and educational expenditure, but not MCAS test score. This finding sort of follows conventional wisdom, that people pay a premium for property where there is a higher focus on education, though usually we expect a more direct relationship with test scores.

<div align="center">

Table 9

Town Model

	Coefficient	Standard Error	p>|t|
constant	207.4	3.102	0.000
perlimit	-0.322	0.123	0.014
percapy	1.093	0.128	0.000
voc	-4.029	1.214	0.003
R²	0.8905	0.8788	

</div>

Following Bolon's main concern, the best model from town data is quite familiar. I show its coefficients in Table 9. It is my five factor model less "exam" and "bosnoex," which have been aggregated into a single Boston observation.

I contend that the highest scoring students are exactly who we would expect them to be, those from highest income places, modified by those student characteristics that we expect to generate lower scores, such as vocational curriculum and limited ability in English. Boston fits right in. The most "influential" observation, by far, is Marblehead. Without that observation, the adjusted R^2 = .9221, the coefficient for perlimit is -2.73 (p = .011), and the coefficient for percapy is 1.247 (p = .000). The coefficient for voc is close to that in Figure 9. Thus the general effect of limited English ability (as well as income) on MCAS score is quite a bit larger than Bolon suggests, or that my model on all towns suggests. If we are going to delete data, let us do so to improve the generalizability of the results. When we do that, we emphasize the importance of the points I have made (that "voc" and "perlimit" need to be in any model explaining test score in these data).

Final Remarks

There is nothing wrong with analyzing a set of data and concluding "Nothing new here, looks like these relationships usually do." Unfortunately, such findings are seldom published. So we systematically lose confirming information, or broadening information (that relationships found elsewhere apply here) that Bolon could have provided. There *is* something wrong with saying one has analyzed data and found that "community income swamped the influence of the other social and school factors examined" when that conclusion is drawn from failing to utilize the information at hand, and dropping two-fifths of the observations. Bolon concludes: "Large uncertainties in residuals of school-averaged scores, after subtracting predictions based on community income, tend to make the scores ineffective for rating performance of schools." What are these "uncertainties?" Why should the state care about the residual from Bolon's model? That his model produces residuals that do not correlate perfectly from year to year implies nothing about the effectiveness of rating schools that, as shown here, do pretty much line up the same every time we look.

He goes on: "Large uncertainties in year-to-year score changes tend to make the score changes ineffective for measuring performance trends." Does he mean "large *variation*?" Variation may not be uncertainty. His attempt to define "large" from published student reliability data is innovative and bold but, though I have not dwelled on it here, not convincing in a study of school averages. Wouldn't "trends" be measured by consistency of direction? Unless Bolon can tell us that there are trends which these data fail to pick up, what he is saying is that "trend" studies should account for random variation, that to establish a "trend" means to exceed variation from changes in the students and tests and random factors from year to year. Well yes, of course. Is he saying no school could meet this criterion? Quite the contrary, he says that too many do. Where in all other respects, when Bolon finds a result inexplicable by chance, he accepts it as "significant," in his "trend" study he rejects the results just because they appear to be inexplicable by chance.

When he concludes that "tenth-grade MCAS mathematics tests mainly appear to provide a complex and expensive way to estimate community income," Bolon is ignoring the fact that MCAS tests are designed to measure individual achievement. If they do that job, then Bolon will find that individuals who do well traditionally are found with others who do well. His finding of a correlation of score with income does not argue that the test fails to measure understanding of mathematics. Does Bolon not believe that more of the "best" academic students, in general, attend Boston Latin, or Brookline, Newton, Lexington and a few other elite schools, than other schools? Does he not believe that the three Boston exam schools "cream" the best students from other schools, leaving these other schools the lowest scoring in the area?

What is *wrong* with finding that high scores in attributes that produce wealth are correlated with the wealth they produce? This is one reason my parents moved to Brookline, and surely one reason Bolon lives there, too. Why does Bolon not want to emphasize

the more important findings that students who have difficulty with English, or with an academic environment, do not do well on this academic test? I do not know why the Massachusetts Department of Education wants to impose this test on all students, or why it wants to deny a high school graduation to those who fail it. However, nothing in Bolon's article argues against them. Bolon is free to disagree with the state (as would I), but saying that there is a high correlation between test score and community income, not news in the twenty-first century, does not support his position.

Notes

[1] Bolon would disagree. He questions their reliability in Tables 2-1 and 2-2 based on year to year changes in school averages. I discuss reliability to some extent below. I agree that Bolon raises good questions, but I disagree that Bolon can criticize test reliability, as understood by test makers, in this manner.

[2] Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York, Equality of Educational Opportunity (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).

[3] I see no rationale for utilizing land value as an independent variable predicting test scores. By the middle 1970s, urban planners had determined that property values reflect school scores. People want higher scoring friends for their children. No one has shown the research that says people act this way, or their behavior, to be wrong. Bolon has the direction of causality reversed even as a hypothesis.

[4] Bolon, last paragraph of Section 1, Part B.

[5] The probabilities are from the t- distribution applied to the difference in scores divided by the joint standard deviation (the square root of the sum of the variances). However the Boston-other comparison is two tail, because we have no hypothesis about which schools would score higher, whereas the exam-not exam comparison is one-tail, where a prior hypothesis about the direction of the difference is clear.

[6] Third paragraph of Section 1, Part B. He also says "I choose to include such schools in these studies while noting their special character." In no part of his analysis is their "special character" noted, although it could have been.

[7] First quote from Bolon p. 4, second from p. 6.

[8] I was plaintiffs' statistical expert in the "Comparable Worth" litigation in the state of Washington in 1983. I weighted the salary structure of jobs by the number of employees per job, because salaries in larger jobs were in fact set more carefully than those in smaller jobs. The larger jobs contained more information about whether gender appeared to be a factor in determining salary, so weighting was appropriate.

[9] As Bolon says, "the data generally available for test score research fail to capture much of the critical information needed to understand development of cognitive abilities and educational achievement in the settings of public schools." Section 1, Part D, first paragraph.

[10] Craig Bolon kindly sent his data to me. I did not examine each item, but I did assure that the correlation between his data and mine, for each variable, was 1.0000. It is conceivable, but unsettling if true, that his program Statistica, produces different results from mine, Stata.

[11] Bolon, second paragraph after Table 2-13.

[12] Bolon equates R^2 with "accuracy." "In an attempt to improve accuracy of the model in Table 2-14, schools with residuals from the two-factor model for 1999 that were greater than two standard deviations were dropped." (at Table 2-18). Influence, not residual size, can be used to delete variables to increase the generalizability of results. See below. All that is accomplished by deleting large residuals is configuring the data to support the model and report higher than real R^2. For example: "Community income has been found strongly correlated with tenth-grade MCAS mathematics test scores and associated more than 80 percent of the variance in school averaged 1999 scores for a sample of Boston-area communities." (Bolon Section 3, Part B, "Conclusions.") From Table 7, below, one can see that he would have had to report "associated more than 47 percent of the variance . . . " from his original sample.

[13] When my 5-factor model is run on 2000 data, all variables have probability .006 or less. On 1998 data, the "voc" variable has a probability .064, perlimit .041, and all other variables .002 or less. Would Bolon delete "voc" from only the 1998 model for its "insignificance?" I hope not.

[14] Bolon, just before Section 2, Part C, "Observations."

[15] Section 2, part D, opening sentences.

About the Author

Stephan Michelson

Stephan Michelson is co-founder and President of Longbranch Research Associates (LRA). Formed in Washington, DC in 1979, now with offices in Maryland, North Carolina and Oregon, LRA provides statistical analysis in litigation. Dr. Michelson has been on the faculties of Reed College and the Harvard Graduate School of Education, and had minor associations with Stanford University (summer) and University of Maryland (Adjunct Professor). He has been on the staffs of The Brookings Institution, Center for Law and Education, Center for Community Economic Development and The Urban Institute (Senior Fellow).

130

Adrián Acosta (México)
Universidad de Guadalajara
adrianacosta@compuserve.com

Teresa Bracho (México)
Centro de Investigación y Docencia Económica-CIDE
bracho dis1.cidc.mx

Ursula Casanova (U.S.A.)
Arizona State University
casanova@asu.edu

Erwin Epstein (U.S.A.)
Loyola University of Chicago
Eepstein@luc.edu

Rollin Kent (México)
Departamento de Investigación Educativa-
DIE/CINVESTAV
rkent@gemtel.com.mx   kentr@data.net.mx

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

J. Félix Angulo Rasco (Spain)
Universidad de Cádiz
felix.angulo@uca.es

Alejandro Canales (México)
Universidad Nacional Autónoma de México
canalesa@servidor.unam.mx

José Contreras Domingo
Universitat de Barcelona
Jose.Contreras@doe.d5.ub.es

Josué González (U.S.A.)
Arizona State University
josue@asu.edu

María Beatriz Luce (Brazil)
Universidad Federal de Rio Grande do Sul-UFRGS
lucemb@orion.ufrgs.br

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gscisuela.cdu

Education Policy Analysis Archives

Confusing the Messenger with the Message:
A Response to Bolon

Victor L. Willson
Texas A&M University

Thomas Kellow
University of Houston

Abstract
The conclusions by Bolon (2001) based on the relationship between per capita income and school mean grade 10 mathematics scores in Massachusetts and on instability in year-to-year mean school scores are criticized by us. Our concerns focus on the uninterpretable covariation of economic condition with test performance and the limitations in interpreting cross-time variability. We agree with Bolon's conclusions but consider the methodology employed inadequate to support them. We suggest alternative requirements and discuss our own previous efforts in this area.

In an analysis of the Massachusetts graduation examination, Bolon (2001) examined the aggregate grade 10 mathematics test scores for 47 high schools and the demographic characteristics of the communities in which they were situated. From several data analyses, Bolon determined that since the best single predictor of mean high school score was community per capita income,

> "The state is treating scores and ratings as though they were precise educational measures of high significance. A review of thenth-grade mathematics test scores from academic high schools in metropolitan Boston showed that statistically they are not."

Further, when removing the variability due to per capita income,

> "Large uncertainties in residuals of school-averaged scores, after subtracting predictions based on community income, tend to make the scores ineffective for rating performance of schools. Large uncertainties in year-to-year score changes tend to make the score changes ineffective for measureing performance trends."

132

While we agree with Bolon's concerns, on the whole, we find little support in the evidence he presents to support them. Our discussion below details our concerns.

Predicting aggregate test scores

One of the problems with regression analysis is that without reasonable theoretical support, all sorts of predictors can be found that produce high correlation. In examining aggregate scores, such as high school test means, it is no secret that for many decades, as Bolon himself pointed out (Bolon, 2000), achievement has been associated with socioeconomic conditions in communities. In earlier eras, when school spending was much more unequal, these differences were more indicative of opportunity to learn for students. In a judicial climate that has tended to minimize, although not eliminate such disparities, it is much less persuasive, although it remains an important area for study.

The difficulty with using a community aggregate measure as a predictor is that it is a surrogate for many other indicators, some of which are absurd at face value but interpretable. Variables such as driver's-license passing rate or per capita champagne consumption may predict student achievement as well as community per capita income. We can construct meaningful arguments why they might. For none is the test invalidated using accepted standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

In other areas of research such aggregation has produced fundamentally misleading conclusions. For example, the literature on intelligence and income is directly parallel to the discussion here. White (1982) demonstrated the difference between using an aggregate measure of SES (school or community) and individual measure in relating SES to intellectual functioning. Since Bolon used school as his unit of analysis, he eliminated proximate measures more appropriate to his analysis. The school-level variables Bolon eliminated are more appropriate than community per capita income on this basis if in fact they were school-based and not district-based. Measures such as free and reduced lunch (FRL) are better indicators for elementary school than for secondary school analyses, however, because of social undesirability of either participating or reporting among secondary students, who tend to have independent means for buying lunches.

The principle of proximity in selecting variables should be carefully considered and invoked. Mixing levels of analysis produces uninterpretable results, as hierarchical linear modeling advocates have pointed out. Bolon erred in this way, we argue.

Test validity: AERA/APA/NCME Standards

The Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999) list 24 points related to validity. We will review those we believe to be relevant to Bolon's argument and attempt to show that his representation is irrelevant to any of them. Standard 1.1 requires a rationale for each recommended interpretation and use of test scores with a summary of evidence and theory. Standard 1.6 requires content validity procedures to be described and justified. While we do not pretend to know in detail the Massachusetts tests, we have a great deal of familiarity with those in our own state, and with the arguments focused on such high stakes tests. The foremost rationale presented in all such state testing programs is content of the state curricula or guidelines.

Challenges to content validity have been consistently thrown out by courts, including a recent case here in Texas (Mehrens, 2000, citing *GI Forum et al vs. TEA et al.*, CA No. SA-97-1278-EP, U.S. District Court, Western District of Texas, San Antonio, TX). Mehrens invoked the 1985 Standards to review the Texas statewide assessment in a process we follow more briefly here.

The congruence of test content with intended instruction is a central focus of test development. Nothing about test content appropriateness was evident in the analysis of income prediction of performance by Bolon. A more comprehensive and focused analysis might ask if schools in lower income communities do not adhere to the state guidelines, or if their teachers are unprepared to teach the mathematics required, or suitable textbooks are not available, so that students do not have an opportunity to learn. These representations might make a case for the relevance of income in dismissing the mathematics test as a precise educational measure of high significance.

The per capita income disparities in our state are much greater than those shown by Bolon. Our experience with our own Texas Assessment of Academic Skills (TAAS) at all levels and content areas has convinced us that income inequities, while important, are not the most useful explanatory variable in school performance. With much larger databases available to us, such as multiyear summaries of all schools in Texas by grade, we see much greater variation in school performance than is shown in the 47 schools Bolon selected. While the correlation is much weaker than the .9+ Bolon presented, nevertheless it is substantial and meaningful. When looking at scatterplots of performance, however, we are struck by the existence of very high-poverty community schools that manage to score very high on the TAAS. For example, Fig. 1 shows the scatter for about 3000 schools with 3rd grade classrooms of TAAS reading and percent economically disadvantaged students (school level measurement), what we would call a surrogate for per capita income. What is of interest is the top of the graph, and the many schools that perform in a manner the state defines as excellent. The correlation data are reported in Table 1 (the approximately 800 schools not reporting economic disadvantage had somewhat higher TAAS scores than those are that did report).

133

Percentage Economic Disadvantaged
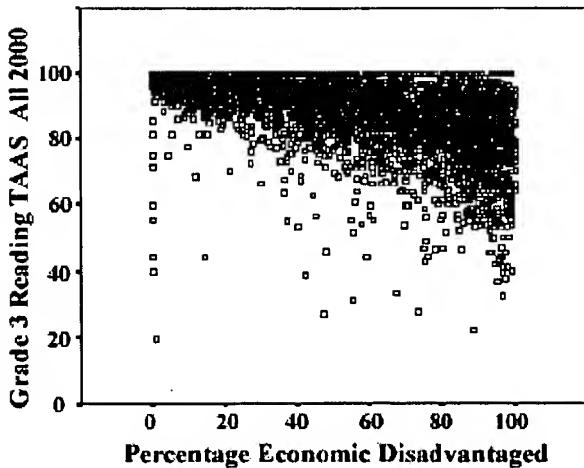
Correlations

		GR 3 READING TAAS ALL 2000	PERCENTAGE ECON DISADV
GR 3 READING TAAS ALL 2000	Pearson Correlation	1.000	-.473**
	Sig. (2-tailed)	.	.000
	N	3671	2893
PERCENTAGE ECON DISADV	Pearson Correlation	-.473**	1.000
	Sig. (2-tailed)	.000	.
	N	2893	3089

**. Correlation is significant at the 0.01 level (2-tailed).

Group Statistics

	INCOMEGP	N	Mean	Std. Deviation	Std. Error Mean
GR 3 READING TAAS ALL 2000	.00	778	88.8530	8.6122	.3088
	1.00	2893	86.8944	11.6742	.2170

Per capita income is an uninterpretable predictor; its relatedness, or not, to school achievement tells us nothing about the stakes being tested, high or not. It fails the theory criterion of standard 1.1.

Instructional effects

While income is related to achievement, whether in Boston or Texas, the central issue is what students enter a school year knowing, what the school teaches them, and what part of the cotent taught is assessed by the end of year test. Standard 1.15 is most relevant: "When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided." Per capita income does not provide any insight into this, nor does, unfortunately, year-to-year change score.

We are unaware of any state that has actually conducted an instructional effect study with pilot versions of its tests to examine the sensitivity of their high stakes tests to instruction. The first author was a member of a committee formed by the legislature of the state of Texas to recommend the structure of the current accountability system (College of Education, Texas A&M University, LBJ School of Public Affairs, The University of Texas at Austing, College of Business, University of Houston, 1993). In the course of committee discussion, the suggestion was raised by the first author that only with some form of pre-post within year assessment at the student level would there be even minimal evidence for instructional change. This suggestion was ultimately rejected by politicians as too costly to consider. Instead, year-to-year student (and school) change was later made into amethodologically suspect statistic, the Texas Learning Index. Bolon has made the same error in considering longitudinal change within test. The alternative

explanations for yearly change negate any interpretation about large uncertainties. Student composition, student mobility, curricular emphases, teacher stability, administrative upheavals, and historical internal validity threats all may explain the variation in a school. Unless and until those are explored and discounted, Bolon's analysis does not support any particular validity threat to the test. We agree that schools, before being held accountable, must be examined carefully for the alternatives listed above. Year-to-year comparisons are inherently flawed due to internal validity threats; the connection between instruction and student performance is weak. It is only because of unwillingness to investigate the actual productivity of the school that a year-to-year comparison is made.

Content limitations

Another major limitation in any interpretation of either static (between school) or dynamic (within school) variation in performance lies in the test items and the sampling of the curriculum. Most high stakes tests are too brief to represent the curriculum adequately. Bolon does not discuss the characteristics of the 10th grade mathematics test. Our experience with the exit level math exam in Texas is that it is unrelated to the content studied in the last 2-3 years (typically grade 8 arithmetic and pre-algebra), and while possessing reasonable internal consistency (.90+), it is too brief to span the domain with only about 40 items. As Mehrens (2000) pointed out for the Texas graduation test, states such as Massachusetts conduct the technical aspects adequately. The standards (either for 1985 or for 1999) will be met. Nevertheless, although important concepts are sampled, the tests are brief, certainly briefer than one would wish to generate a score representing 10 or 11 years' schooling.

The Texas released 10th grade mathematics examination (Texas Education Agency, 2000) has 40 items. From a review of the content, it appears that at best only one or two assess topics not covered in grades 8 or below, while one item (19) appears to be a spatial rotation task more appropriate to an intelligence test. The inadequacy of such a test to evaluate 9 or 10 years' mathematics learning is, if not self-evident, at least empirically testable. One can conceive of various research investigations involving interview with teachers and students and performance demonstrations by students on the full range of TEKS objectives to evaluate how well a short form such as the TAAS estimates actual mathematics declarative and procedural knowledge. In the 1993 discussions in Texas cited above, the introductory letter by Charles Miller (1993) made clear that the committee proposed to eliminate the 10th grade test in favor of specific grade 10 subject matter tests such as Algebra and Biology. While there was an obvious concern for creating a set of hurdles, the committee's recommendation was based on testing students over content more proximate to their instruction.

Year-to-year stability

Table 1 presents correlations within and across year for grades 3-5 for 1999-2000

Table 1
Correlations within and across year for grades 3-5 for 1999-2000

Correlations

		GR 3 READING TAAS ALL 2000	GR 3 READING TAAS ALL 1999	GR 4 READING TAAS ALL 2000	GR 4 READING TAAS ALL 1999	GR 5 READING TAAS ALL 2000	GR 5 READING TAAS ALL 1999
GR 3 READING TAAS ALL 2000	Pearson Correlation	1.000	.649**	.674**	.567**	.668**	.575**
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000
	N	3671	3543	3435	3337	2969	2877
GR 3 READING TAAS ALL 1999	Pearson Correlation	.649**	1.000	.658**	.647**	.579**	.641**
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000
	N	3543	3543	3326	3316	2862	2858
GR 4 READING TAAS ALL 2000	Pearson Correlation	.674**	.658**	1.000	.636**	.699**	.608**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000
	N	3435	3326	3435	3316	2948	2854
GR 4 READING TAAS ALL 1999	Pearson Correlation	.567**	.647**	.636**	1.000	.694**	.699**
	Sig. (2-tailed)	.000	.000	.000	.	.000	.000
	N	3337	3316	3316	3337	2858	2861
GR 5 READING TAAS ALL 2000	Pearson Correlation	.668**	.579**	.699**	.694**	1.000	.660**
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000
	N	2969	2862	2948	2858	2969	2842
GR 5 READING TAAS ALL 1999	Pearson Correlation	.575**	.641**	.608**	.699**	.660**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.
	N	2877	2858	2854	2861	2842	2877

**. Correlation is significant at the 0.01 level (2-tailed).

Note that within year, cross-grade correlations are higher than between-year, within grade correlations or between-year, cross-grade correlations, generally. The cohort effect, however (grade 3 in 1999 to 4 in 2000, grade 4 in 1999 to grade 5 in 2000) appears supported since these correlations are higher than the cross-year, within-grade correlations for grades 3-5. That at least appears consistent with what might be expected: between-student correlations are lower than within-student correlations, however attenuated they might be for school averages. If these correlations are to be used as part of a school-level assessment, however, they appear woefully inadequate psychometrically.

A different approach can be taken by treating the within-year cross-grade TAAS scores as scale variables. Then coefficient alpha for 1999 is.8539 and for 2000 is .8637. If one is evaluating schools, these are reasonably good values.

Table 2 presents change correlations. While cautions abound about interpreting change scores, if change is the currency to be used, descriptive and correlational characteristics must be considered. Obviously, the change measures that include the same score, such as D33 with D43, are inflated by the self-covariation. The other correlations, that are not self-inflated, are generally positive but modest, almost all in the .1 to .2 range. The D43 and D54 correlation of .165, for example, supports a conclusion that schools' cohorts improve together (or fall behind together).

Table 2
Correlations for yearly change within and between for grade 3-5 and changes for grades 3 to 4 and 4 to 5 for Texas schools 1999-2000

Correlations

		D33	D44	D55	D43	D54
D33	Pearson Correlation	.649	.122**	.179**	.569**	.234**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	3543	3302	2824	3326	2838
D44	Pearson Correlation	.122**	.636	.113**	.436**	.541**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	3302	3316	2821	3302	2848
D55	Pearson Correlation	.179**	.113**	.660	.186**	.504**
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	2824	2821	2842	2813	2831
D43	Pearson Correlation	.569**	.436**	.186**	1.000	.165**
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	3326	3302	2813	3326	2834
D54	Pearson Correlation	.234**	.541**	.504**	.165**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.
	N	2838	2848	2831	2834	2858

**. Correlation is significant at the 0.01 level (2-tailed).

Note: D33, D44, and D55 are the stability coefficients in Table 1.

Table 3 presents correlations between change measures and available school characteristics such as size of school (ENROLL 2000), percent of school on free-and-reduced lunch, percent of school with Limited English Proficient (LEP) students, and percentage of the ethnic groups that in Texas are the focus of civil rights enforcement (African-American and Hispanic). LEP students were exempted from the TAAS in these years, so high performing schools with high LEP percentages can be based on very small samples of white students. The correlations are of interest insofar as they provide different ways to examine school performance at the individual school level rather than using the block method employed by Bolon.

Conclusions

Our concern with the Bolon (2001) study was that it focused on a relationship between school performance on a high stakes test and community wealth that is not informative about the characteristics of the test. The emphasis on wealth and its relationship to schooling has been highlighted in educational thought since Coleman's (Coleman, Campbell, Hobson, McPortland, Mood, Weinfeld, & York, 1966) conclusions about the efficacy of schooling. These aggregated analyses have not, we contend, illuminated much about why schools succeed or fail. Studies about schools focusing on leadership and its relationship to school performance, for example, provide meaningful, interpretable, and actionable conclusions for school level policy. Unfortunately, barring exchanges of cash between communities, Bolon's work does not.

Table 3

Correlations between TAAS 1999-2000 grade changes and selected school characteristics

Correlations

		D43	D54	D55	D44	D33	ENROLL 2000	% ECON DISA DV	% LIMIT ENGL PROF	% AF-AM	% HISP	% WHITE
D43	Pearson r	1.000	.165**	.186**	.436**	.569**	-.063**	.139**	.012	.189**	.037	-.127**
	Sig.	.	.000	.000	.000	.000	.001	.000	.540	.000	.055	.000
	N	3326	2834	2813	3302	3326	2764	2764	2550	2446	2756	2715
D54	Pearson r	.165**	1.000	.504**	.541**	.234**	-.095**	.028	-.114**	.035	-.029	.008
	Sig.	.000	.	.000	.000	.000	.000	.142	.000	.087	.129	.675
	N	2834	2858	2831	2848	2838	2755	2755	2540	2432	2747	2704
D55	Pearson r	.186**	.504**	1.000	.113**	.179**	-.054**	.173**	.077**	.050*	.128**	-.142**
	Sig.	.000	.000	.	.000	.000	.004	.000	.000	.014	.000	.000
	N	2813	2831	2842	2821	2824	2740	2740	2530	2421	2732	2691
D44	Pearson r	.436**	.541**	.113**	1.000	.122**	-.054**	.134**	.011	.078**	.067**	-.098**
	Sig.	.000	.000	.000	.	.000	.005	.000	.568	.000	.000	.000
	N	3302	2848	2821	3316	3302	2760	2760	2546	2442	2752	2709
D33	Pearson r	.569**	.234**	.179**	.122**	1.000	-.081**	.045*	-.030	.090**	-.014	-.026
	Sig.	.000	.000	.000	.000	.	.000	.017	.131	.000	.474	.178
	N	3326	2838	2824	3302	3543	2785	2785	2568	2463	2777	2736
ENROLLMENT 2000	Pearson r	-.063**	-.095**	-.054**	-.054**	-.081**	1.000	.032	.256**	-.083**	.180**	-.239**
	Sig.	.001	.000	.004	.005	.000	.	.079	.000	.000	.000	.000
	N	2764	2755	2740	2760	2785	3089	3089	2743	2709	3064	3025
PERCENTAGE ECON DISADV	Pearson r	.139**	.028	.173**	.134**	.045*	.032	1.000	.600**	.265**	.692**	-.791**
	Sig.	.000	.142	.000	.000	.017	.079	.	.000	.000	.000	.000
	N	2764	2755	2740	2760	2785	3089	3089	2743	2709	3064	3025
PERCENTAGE LIMIT ENGL PROF	Pearson r	.012	-.114**	.077**	.011	-.030	.256**	.600**	1.000	-.160**	.762**	-.666**
	Sig.	.540	.000	.000	.568	.131	.000	.000	.	.000	.000	.000
	N	2550	2540	2530	2546	2568	2743	2743	2743	2431	2743	2694
PERCENTAGE AFRIC-A	Pearson r	.189**	.035	.050*	.078**	.090**	-.083**	.265**	-.160**	1.000	-.313**	-.383**
	Sig.	.000	.087	.014	.000	.000	.000	.000	.000	.	.000	.000
	N	2446	2432	2421	2442	2463	2709	2709	2431	2709	2692	2664
PERCENTAGE HISPANI	Pearson r	.037	-.029	.128**	.067**	-.014	.180**	.692**	.762**	-.313**	1.000	-.789**
	Sig.	.055	.129	.000	.000	.474	.000	.000	.000	.000	.	.000
	N	2756	2747	2732	2752	2777	3064	3064	2743	2692	3064	3004
PERCENTAGE WHITE	Pearson r	-.127**	.008	-.142**	-.098**	-.026	-.239**	-.791**	-.666**	-.383**	-.789**	1.000
	Sig.	.000	.675	.000	.000	.178	.000	.000	.000	.000	.000	.
	N	2715	2704	2691	2709	2736	3025	3025	2694	2664	3004	3025

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Note: D33=grade 3 change from 1999 to 2000; D43 = change from grade 3 in 1999 to grade 4 in 2000, etc. Lunch (% ECON DISADV), and percentage of targeted minority groups in Texas, African- Americans (% AF AM) and Hispanics (% HISP), as well as majority whites (% WHITE).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Bolon, C. (2000). School-based standard testing. *Education Policy Analysis Archives* 8(23), May, 2000, available at http://epaa.asu.edu/epaa/v8n23.

Bolon, c. (2001). Significance of test-based ratings for metropolitan Boston schools. *Education Policy Analysis Archives* 9(42). October 16, 2001, available at http://epaa.asu.edu/epaa/v9n42.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPortland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity. Washington, DC: U.S. Government Pringing Office, 1966.

College of Education, Texas A&M University, LBJ School of Public Affairs, The University of Texas at Austing, College of Business, University of Houston, (1993). *A New Accountability System for Texas Public Schools, Vol. 1.* Austin: Educational Economic Policy Center, The University of Texas at Austin.

Mehrens, W. A.(2000). Defending a state graduation test: *GI Forum et al. vs. Texas Education Agency.* Measurement perspectives from an external evaluator. *Applied Measurement in Education, 13*(4), pp. 387-401.

Miller, C. (1993). Introductory letter. In College of Education, Texas A&M University, LBJ School of Public Affairs, The University of Texas at Austing, College of Business, University of Houston, *A New Accountability System for Texas Public Schools, Vol. 1.* Austin: Educational Economic Policy Center, The University of Texas at Austin.

Texas Assessment of Academic Skills Exit Level . Austin: Texas Education Agency, February 2000. Available at http://www.tea.state.tx.us/student.assessment/resources/release/taas/release01/xl.pdf

White, Karl R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461-81.

About the Authors

Victor L. Willson
Professor
Department of Educational Psychology
Texas A&M University
College Station, Texas

Email: v-willson@tamu.edu

Victor L. Willson is Professor of Educational Psychology and Teaching, Learning and Culture, and Director of the Cognition and Instructional Technologies Laboratory, Texas A&M University. His current research interests focus on the intersections of psychometrics, children's reading development, and individual differences in cognitive measurement.

J. Thomas Kellow
Visiting Assistant Professor
College of Education
University of Houston

Email: tkellow@pdq.net

J. Thomas Kellow is currently Visiting Assistant Professor in the College of Education at University of Houston. He received his Ph.D from Texas A&M in Educational Psychology, and has research interests in high-stakes testing, applied statistics, and disability studies with an emphasis on the community integration of persons with mental retardation.

139

Universidad de A Coruña
jurjo@udc.es

University of California, Los Angeles
torres@gseisucla.edu

140

Education Policy Analysis Archives

Response to Michelson and to Willson and Kellow

Craig Bolon
Planwright Systems Corporation (USA)

Abstract
The criticisms and points made by both Michelson (http://epaa.asu.edu/epaa/v10n8/) and Willson and Kellow (http://epaa.asu.edu/epaa/v10n9/) in response to my article "Significance of Test-based Ratings" (http://epaa.asu.edu/epaa/v9n42/) are here addressed.

Michelson's Complaints

Michelson's critique of "Significance of Test-based Ratings" rides herd on some fine points but misses main themes of the article. The article's data include test scores for only 47 schools. As experts have warned, such a small behavioral data set can typically provide stable coefficients for only one or two independent variables. The work leading to the article aimed to see if one or two strong variables could be found for this limited data set. As it happened, a dominant variable was found: community income.

In seeking an expanded analysis, Michelson overloads the observations with independent variables, adding some with no evidence for underlying quality. While his approach associates more variance with a larger set of variables, he does not provide stepwise or combinatorial analysis for the incremental association of variance or conduct a sensitivity study to explore the likelihood that his results may be an artifact. Despite an undiscriminating approach, community income remains the strongest factor.

The article, by contrast, emphasizes robust results obtained from accurate, traceable data and parsimonious models. It employs cross-validation, sensitivity analysis and combinatorial analysis. Weighting is introduced to construct models that will not be destabilized by smaller schools contributing to the data set.

Apparently unsatisfied with his alternative, since community income is still the strongest factor, Michelson proceeds with polemics centered around the notion that the article really has no news anyway, since (somehow) everybody knows that high test scores go along with high incomes. Maybe everyone in his circle does, but many people I encounter are surprised; they wonder why this should be so.

It is known, if not well known, that by the early 1920s labor unions mounted protests against the social injustice of using IQ scores to place students in academic "tracks." They had found out that high IQ scores were strongly

141

associated with high family incomes. It is also known, if not well known, that by the mid-1950s the Educational Testing Service had found a regular progression of their average SAT scores with average reported family incomes. But these results are for "aptitude" tests.

There have been limited published studies about the associations of social and school factors with state "accountability" test scores. Will such a test be similar in social correlates to "aptitude" tests, or will it be different? Massachusetts is a useful laboratory for such a study. It did not begin "accountability" testing until 1998. It then created what is generally regarded as a state-of-the-art program.

The Massachusetts graduation requirement will not directly affect any student until 2003. Although Massachusetts has some communities with a history of aggressive testing, such as Worcester, before the school year ending in 2000 most schools made light to moderate responses to MCAS tests. These circumstances help provide a good baseline.

Problems for such studies are the rare availability of reliable, personal social-factors data and the limited social-factors data that one can clearly associate with individual schools. Most data collected and reported by schools either count disadvantaged populations or count eligibility for free or reduced-price lunch. This produces two common outcomes. One is correlations found for test scores with population categories. The other is correlations found between test scores and poverty, since poverty or near-poverty income is the qualification for free or reduced-price lunch.

Beginning with Massachusetts school profiles, I also found those correlations. But being familiar with the communities for which I had data, I decided to look at residuals. To me the residuals seemed to show a pattern—high score-residuals in high-income communities and vice-versa. Data for disadvantaged populations, poverty and other school categories did not appear to tell the whole story, so I sought income data for school populations.

It turned out that the only generally available data were from the 1989 U.S. Census of Population and Housing. Somewhat to my surprise, per-capita community income from a decade prior to the test scores proved to be a strong and robust factor. The article recounts the modeling of data in the sequence it occurred.

The major theme, which Michelson does seem to understand, is that income appears to matter at all ordinary levels, not just at the threshold of poverty. Part of this may be self-fulfilling prophecy, when test scores are used to grant or deny advancement, but there is probably more. I don't accept Michelson's guesses about the phenomenon. I have different hypotheses but won't trust those either without evidence.

Another theme, which Michelson seems to ignore, is that community income, as distinct from family income, may have a powerful effect. That is merely a suggestion in the article, of course. It would take a study of individuals to differentiate the influences.

A fortuitous circumstance for this study was the pattern of New England cities and towns, which form legal boundaries around small, diverse clusters of population. That can also be found elsewhere in the U.S., such as near Philadelphia or Cleveland, but in the more recently settled areas it is rare. Instead, a large city has usually been allowed to swallow up many neighbors, and the remaining suburbs do not have the diversity of the urban neighborhoods. Social data collected within city boundaries can be very difficult to reaggregate, as happens in the City of Boston.

The following paragraphs respond to Michelson's observations item by item:

(1) Michelson first complains about what he calls a "non-sequitur." The article's abstract says, "The state [of Massachusetts] is treating scores and ratings as though they were precise educational measures of high significance." And indeed it is doing just that. As the article later points out, getting 23 instead of 24 correct answers on a tenth-grade mathematics test can be enough for Massachusetts to deny a student a high-school diploma. Such a small difference can produce a huge effect, since high-school graduation has great influence on lifetime income.

The article's abstract goes on to say that a "review of tenth-grade mathematics test scores...showed that statistically [Massachusetts scores and ratings] are not [precise educational measures of high significance]." And indeed the scores are not precise. As the article shows, the variability of score averages is so large that at least several years will be needed to see whether there are definite trends for most schools. As the article also shows, the educational significance of the scores is highly questionable. As with "aptitude" tests, scores closely track income levels. Once predictions based on income have been subtracted, few schools can be distinguished. There is little to indicate that these scores may measure what schools achieve, as contrasted with what social advantages or disadvantages students bring to schools from their backgrounds.

(2) Michelson complains the article "has not said what success on these tests is supposed to imply." That's not my job; it's the job of an agency in charge of the tests. What the article says is that no studies "have shown that MCAS test scores have practical significance, in the sense of predicting success in adult activities to any greater degree than could be done with knowledge of student backgrounds."

It should be a responsibility of government, when using tests so as to cause drastic life consequences for young people under its care, to demonstrate objectively that its tests accurately and fairly measure skills of critical and lasting importance. Massachusetts has failed to do this. It merely says its tests are similar to other tests, which also lack practical validation. A physician acting in such a cavalier way toward a patient would be at risk of fines or jail.

The topic of the article, however, is school ratings. We maintain public schools to equip all young people with skills and knowledge essential to support themselves and to carry out civic responsibility. Schools in rich and poor communities alike strive to do this. School ratings that mostly track incomes of communities are unlikely to reflect actual levels of effort or achievement by the schools. The Massachusetts tests lack practical validation, and the ratings based on those tests appear to measure characteristics of communities more than they do those of schools.

(3) Michelson complains about dropping Boston schools from part of the data analysis. That part of the analysis focuses on community income. As the article indicates, given Boston's complex mix of exam schools, magnet schools, district schools, cross-enrollment and busing, it was not possible to determine community income for individual schools. If one wants to expand the data set, it would more fruitful to add other Massachusetts cities and towns than to spend the large amount of effort needed to reaggregate Boston data by schools with any accuracy.

(4) Michelson complains about weighting by number of test takers, but aside from an appeal to prejudice he does not try to explain why an unweighted analysis would be of more use. For example, one way to handle complaint (3) might be to aggregate all of Boston's schools and use citywide per-capita income and other factors. Would it then be useful to treat all of Boston, with a population of 589,141, as equivalent in weight to Winthrop, with a population of 18,303? Despite this grievance, as Michelson later shows, unweighted analysis leads to similar patterns of results. No reviewer of this article raised concerns about weighted models.

(5) Michelson objects to lack of a marker variable for schools with vocational programs in the same facility as academic programs. However, some such schools have only vestigial programs, while other schools enroll large fractions of their students in vocational programs, with year-to-year changes depending on local circumstances. Massachusetts school profiles did not record these programs uniformly when study data were assembled and do not provide program enrollments. A marker variable can be wildly inaccurate. Later Michelson also proposes a marker variable for exam schools. Boston's exam schools are known to vary widely in selectivity, and a marker variable will not account for that.

(6) Michelson's complaint about weighting by numbers of test takers while also including school population as a variable in one of the analysis steps is reasonable; the correlation is quite high. However, I supplied Michelson with a file of all data used for the article. As I saw and he should readily have found, both weighted and unweighted analysis show low significance for this variable. I did not think that putting an additional analysis into the article would add much information, but perhaps a comment about checking unweighted analysis should have appeared.

(7) Michelson's complaint about using only "per pupil expenditure" (regular education) as an estimator for financial support is directed to the wrong party. Massachusetts citizens have protested for years about poor reporting of school spending. As the article says in an appendix, Massachusetts is finally adopting a uniform system of detailed financial reporting. The first published data from the new system will be available some time in 2004.

(8) Michelson makes an attempt to estimate cross-effects of Boston's exam school system in lowering scores of district schools. However, as the article says, there is evidence that many ambitious parents whose offspring who are not accepted to an exam school of their choice send them outside Boston public schools: to parochial schools, other private schools and suburban schools. It would take far more resources than the data for this study provide to investigate cross-enrollment effects accurately.

(9) Michelson objects to the use of unadjusted R^2 to report the variance associated in successive steps of analysis. This would be a reasonable complaint if there were no dominant variable and one needed to account in detail for relative contributions of multiple variables, but that was not the result found in the study. Using adjusted R^2 would intensify the dominance of the community income variable, because adjusted R^2 values obtained when adding more variables to a model are lower than unadjusted values. In his use of adjusted R^2, Michelson does not show that the assumptions of adjustment are actually satisfied for the data.

(10) Under "A Change of Method," Michelson again objects to dropping Boston schools from part of the analysis and seems to miss the point where the change occurs. He states that text following Table 2-10 and a figure in Table 2-13 report an unadjusted R^2 value of .80 for three variables, while Table 2-15, he says, reports a value of .86 for only two variables. This seems to puzzle him; he says he can "see no reason for" it.

Actually, the R^2 value of .86 for two variables is reported in the paragraph preceding Table 2-14. This paragraph begins by saying that analysis starting with that table applies "only to schools outside the City of Boston." A possible cause of the increase in R^2 values is, as the article says, that one cannot accurately determine community income for individual Boston schools. When analyzing mainly schools with unambiguous community income, one is looking at a less noisy data set. The procedure and the reason for it are clearly stated.

143

(11) In his objections to the article's use of unadjusted R^2, Michelson makes much of the difference in R^2 values between the analyses that include the Boston schools and those that don't. Unlike Michelson, the article does not try to compare R^2 between these analyses. They don't use the same data set. Without knowing what is not known about the social factors for individual Boston schools, comparisons like those Michelson suggests will not be meaningful.

(12) Michelson also seems unsure when weighted or unweighted analysis is being used. Weighting by number of test participants is described in the paragraph preceding Table 2-3, the first analysis reported in the article, and was used for that and all following analyses (except, as the article states, those in Figure 2-3). Readers are occasionally reminded that models are being weighted in this way. If Michelson uses unweighted analysis or a different weighting factor, such as school population, then he will get different results.

(13) Michelson's Table 2 and Table 3 offer what he calls an "all schools replication" of analyses in Table 2-14 and Table 2-16 of the article. The two paragraphs preceding Table 2-14 in the article discuss problems of reaggregating data for Boston schools and introduce analyses that consider only the remaining schools, reported in Tables 2-14 through 2-16. An "all schools replication" is not a replication. Since he doesn't use the same data set in his Tables 2 and 3, Michelson gets different results. When he uses the same data set in his Tables 4 and 5, he gets the same results as the article contains.

Michelson notes a concern about excluding only Boston schools in Table 2-14 and Table 2-16, since Quincy, Lynn and Newton also have multiple high schools. The article makes the same observation in its summary analysis, Section 2.D, and presents in Tables 2-20 and 2-21 and in Figures 2-8 through 2-10 results excluding all four of these communities from the data set.

(14) Again, Michelson objects to dropping Boston schools from the data set and proposes a different model into which he introduces marker variables, "Michelson's 5-Factor Model." I considered such an approach during the study but rejected it for the reasons stated under his complaint (5): these marker variables may wildly misrepresent what they claim to identify. In rejecting weighting for his model, Michelson exposes it to instability from smaller schools that are well off the trend lines. He does not explain his reasons for the choice.

Here, from a reader's perspective, Michelson is exploring new ground. He has different techniques for analyzing different sets of data and seems to have different motives. His model ignores conservative recommendations for behavioral data by using too many variables in a final model for the size of the data set. He does not try to overcome the potential problems in this approach with a sensitivity study to explore probable ranges of results from such a model. He does not review any potential weakness of his marker variables. He does not provide readers with stepwise or combinatorial analysis for incremental association of variance, only first and last steps. He does not attempt any cross-validation. He does not explain his rejection of weighting. The results shown in Michelson's Figure 1 may represent a robust pattern, or they may be a statistical artifact.

As previously stated, the article emphasizes robust results obtained from accurate, traceable data and parsimonious models. It employs cross-validation in Table 2-13, sensitivity analysis in the exploration of outliers in and near Tables 2-18 and 2-19, and combinatorial analysis in Table 2-12. Weighting is introduced beginning at Table 2-3 to construct models that will not be destabilized by smaller schools contributing to the data set.

(15) Besides offering no stepwise or combinatorial analysis of his own, Michelson objects to such analysis in the article, shown in Table 2-12 and 2-15 and discussed in Section 2.C. A curious objection, since stepwise and combinatorial analyses are common, helpful approaches to understanding the relative influences of multiple factors. A better objection would have been to call at this point for the use of adjusted R^2, since when the assumptions of adjustment can be satisfied, the adjustment will discount added factors of low significance.

(16) Michelson characterizes the article's discussion of limited English proficiency as "political" and claims the article largely ignores it. Readers can judge for themselves. Section 2.C of the article says, "The factor 'Percent limited English proficiency' was the second strongest influence on predicted test scores." It offers hypotheses for further investigation suggested by this finding. It then goes on to discuss relative significance of factors "Percent African American," "Percent Hispanic / Latino," and "Percent Asian or Pacific Islander." (Unlike Michelson, I prefer longer, more informative factor names to "bosnoex" and other mysterious abbreviations.)

(17) Michelson claims the article ignores what he calls "specification" effects in the evolution of model equations. In fact, several parts of the article address just such effects, and the article as a whole is an extended model development.

In particular, the discussion of Table 2-6 emphasizes how an economic factor captures variance otherwise associated with disadvantaged populations. Table 2-7 is presented to show how a model without the economic factor loads significance onto population categories. The analysis in Table 2-12 shows how different model equations reveal the weakness of one factor. Discussion around Table 2-14 shows how the factor "Per-capita community income (1989)" supplants the significance of the factor "Percent free or reduced price lunch." (Unlike Michelson, I assume readers are familiar with such effects and do not need a lecture.)

144

(18) Arguing about his marker variable for vocational programs, Michelson ignores the lack of data characterizing the programs or the students who enroll in them. It might be that the programs draw many students from low-income families or families who do not speak standard English as a first language. It might be that the programs neglect skills or knowledge being tested by MCAS. There could also be a combination of these factors, or there might be some critical but wholly different factor. Data available for the study were insufficient to address the issue, and the article says so. I believe it is unwise for Michelson to introduce a marker variable without investigating the environment.

(19) Similarly, arguing about his marker variable for Boston exam schools, Michelson relies on personal recollections and anecdotes, but he ignores the complex social characteristics of Boston and the lack of reliable data for estimating cross-enrollment effects, which only begin with the exam schools. A study by the Mumford Center at the University of Albany indicates that parochial and other private schools have such large effects that the incidence of poverty among the households of Boston public school students substantially exceeds the incidence in the city population. As with vocational programs, data available for the study were insufficient, and the article says so. Again, I believe it is unwise for Michelson to introduce a marker variable without investigating the environment.

(20) Caught up in personal anecdotes, Michelson ignores findings about individual communities reported in the article that appear significant. Belmont substantially outscored predictions, while Marblehead scores were considerably lower than predicted. Sensitivity analysis suggests robust results, not artifacts. Neither community is known for extremes in population or education; a review that compares and contrasts them might be of interest and of practical use.

(21) Michelson's objections to the residuals discussion around Figures 2-4 and 2-5 in the article ignores the increase in slope of the line of fit. These figures were placed in sequence so that this effect could be easily seen. Michelson is correct in his observation that annual score averages and score changes are limited predictors—the point made with Figures 2-9 and 2-10 in the article. With Michelson's scatterplot of successive year residuals from "Michelson's 5-Factor Model" in his Figure 3, he might raise a question about the anomalous behavior of Swampscott, which the article called attention to in discussing Table 2-11.

(22) Under "The Town View," Michelson begs the question of the article. He contends that "the highest scoring students are exactly who we would expect them to be, those from the highest income places...." Yes, if you look at the plot of average scores versus income in the abstract of the article, that is what you would expect. But if you hadn't seen the data, would you know? Perhaps, as he seems to suggest, Michelson is privy to inside information. Most of us have to look at the data to find out.

(23) Michelson's "Final Remarks" include the polemics previously mentioned. In these, he contends that "MCAS tests are designed to measure individual achievement" and seems to want to make this an affair of honor. True or false, meaningful or otherwise, that's not the topic of the article, which also seems to have escaped Michelson. Likewise, the article is concerned neither with what Michelson calls "beliefs" about its findings nor (certainly, in an article using statistical inference) with causality.

Summary of Response to Michelson

The topic of the article, "Significance of Test-based Ratings for Metropolitan Boston Schools," is the meaning and usefulness of school ratings that are entirely based on MCAS test scores. The article shows, for the years and tests it reports, that within the variations of test score averages the Massachusetts Department of Education could have produced nearly the same ratings simply by scaling income data from the Department of Revenue. As an appendix to the article notes, Section 11 in Chapter 69 of the Massachusetts General Laws directs the Department of Education to set up a school rating system with a broader approach than it has used so far:

> "The system shall employ a variety of assessment instruments on either a comprehensive or statistically valid sampling basis. Such instruments...shall include consideration of work samples, projects and portfolios, and shall facilitate authentic and direct gauges of student performance."

This provision was written when the state was administering tests on a sampling basis that were inspired by NAEP, which tries to acknowledge a variety of learning orientations. In narrowing its current approach to a single test series, Massachusetts may have emphasized only the cognitive skills sampled by "aptitude" tests. Certainly it fails to honor the spirit of its laws.

Unlike the impression Michelson gives of his outlook, I don't see statistical analysis as a card game, playing to get a high multiple R score while discounting the quality of data. Statistics won't identify causes or distinguish causes from effects. At best one can find robust patterns that justify investigation by other means. Knowing, for example, that community income provides a strong, persistent factor for certain test scores may motivate someone to find out why this happens and might eventually lead to better understanding of how or what to teach or test.

Knowing that factors may have some influence will not help an investigator unless their influence is major. In this vein, someone with a scientific or engineering background will tend to apply a $p<0.05$ criterion as a rough, first cut—

145

the criterion to which Michelson takes such pained exception. That isn't a signal of existential meaning; it's a value judgement. If there isn't much more than a 95 percent chance of significance, then another phenomenon is probably a better object of one's attention. Anyone who tries to chase down all the "Michelson 5-Factor[s]" with surveys or experiments is risking a waste of energy in blind alleys.

Michelson's reported confusions with the article suggest that he would really like to analyze different data sets with different techniques. That's fair, of course, and it might yield some new information. But it really proposes a different article, which Michelson seems to have begun in the guise of a critique.

As "Significance of Test-based Ratings" shows, Massachusetts school ratings, based solely on MCAS test scores, are not precise educational measures of high significance. Score variations are large, and scores appear to reflect heavily the social advantages or disadvantages that students bring to schools from their backgrounds, not necessarily the effectiveness of the schools themselves.

Willson's and Kellow's Concerns

Willson's and Kellow's concerns about "Significance of Test-based Ratings" focus largely on what they say are "theoretical" and standards issues. The article presents data, models and correlations associating state "accountability" test scores used to construct school ratings with social and school characteristics. It does not endorse any "theoretical" framework or invoke any institutional standard of judgement. Willson and Kellow also raise issues about construction and content of tests and discuss data developed by identifying individual students. While these are reasonable subjects of inquiry, they are not the topic of the article.

The article explores the potential significance of school ratings based on "accountability" test scores. The study on which the article was based looked at average test scores and social and school factors for 47 geographically clustered schools to see if one or two strong factors could be found for this limited data set. A dominant factor was found: community incomes. Results show that the test scores track community incomes so closely that it is questionable whether the scores measure efforts or effectiveness of the schools.

Willson and Kellow seem to feel that income inequities should no longer be considered particularly relevant to educational issues, since funding has been equalized. Although both their home state of Texas and mine of Massachusetts have school funding equalization programs, both states also continue to encounter strong disagreements and lawsuits over the issue, for example:

> "In a lawsuit filed against the state of Texas, lower-wealth school districts allege that the state's current funding scheme for education fails to meet the equity and efficiency standards established by a 1994 Texas Supreme Court decision. Plaintiffs in the suit claim that tax exemptions for, among other things, country clubs and sports franchises represent lost property tax revenues that would otherwise be allocated to funding the state's schools. These breaks reduce school funding by an estimated $500 million, according to the plaintiffs. A statewide property tax to correct the state's school funding inequities failed in the last session of the Texas legislature." (from Texas Education Funding, 1998)

> "Residents Advocating Government Equity, or RAGE, sought $50,000 from Barnstable County to continue a private lawsuit, brought in the name of eight Cape schoolchildren to the Massachusetts Supreme Judicial Court. The suit asks the court to intervene and make the state legislature carry out its constitutional duty to adequately finance aid to schools. That includes changing the aid formula to better serve towns like those on the Cape, with relatively high property values but also with high growth rates. Several Cape towns are also owed $14.7 million in back aid." (from Milton, 1999)

Willson and Kellow don't consider recent evidence such as Grissmer, et al., 2000, that income levels may influence test scores independently of school funding. In Figures 2-6 and 2-7 the article "Significance of Test-based Ratings" shows, for the communities studied, that per-capita community income is strongly associated with average test scores while school spending shows very little association with average test scores.

Willson and Kellow make a vague statement that "all sorts of predictors" (other than income) might produce strong correlations with test scores, but they don't offer any evidence. Why not? If strong predictors were so easy to come across, surely they could dredge up a few. Actually, an appendix to the article provided Willson and Kellow with real data to test such a proposition, the social and school factors that the article analyzes. The results are in Figure 1.

146
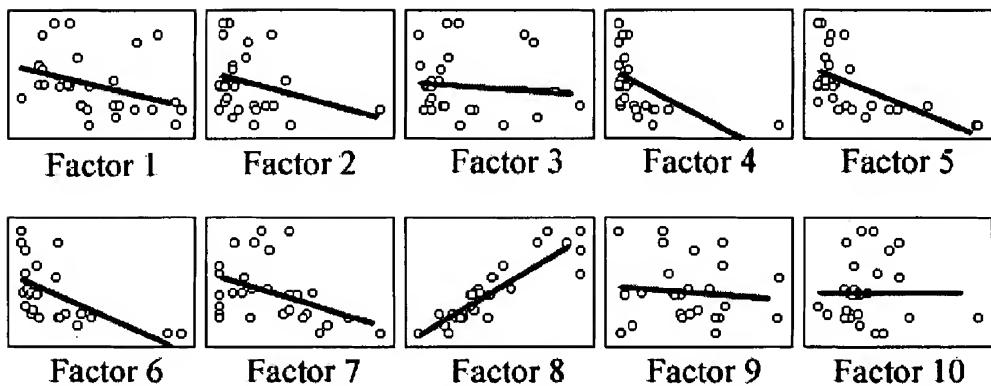
Figure 1. Scatterplots of scores versus factors

The scatterplots in Figure 1 of this response include the schools contributing to the summary analysis of the article, in its Figures 2-8 through 2-10. The ordinate of each scatterplot is the school-averaged test score for 1999. Although several factors have significant correlations, Factor 8 provided dominant and robust association of variance in a multifactor model. In the article, this factor is identified as "Per-capita community income (1989)." Another factor proportional to community income would certainly act as an effective substitute.

Willson and Kellow maintain that using "aggregate measures" may yield "misleading conclusions." Possibly—but from what kind of data are the test-based school rating systems themselves constructed? "Aggregate measures," I believe. Willson and Kellow don't seem to be disturbed that all test-based school rating systems, those they like as well as those they don't, are subject to similar potentials for distortion.

In their objections about "mixing levels of analysis," Willson and Kellow do not seem to have followed the article closely. Preliminary analyses through Table 2-8 use only school-based data. Intermediate steps, introducing per-capita incomes from the census, warn about problems from mixing levels. The summary analysis, Section 2.D, presents in Tables 2-20 and 2-21 and in Figures 2-8 through 2-10 the results from only communities with a single high school. For those schools, there is only one level. The summary analysis has no level mixing.

Despite sensitivities over "mixing levels of analysis," Willson and Kellow sometimes seem to confound issues of evaluating personal test scores in the context of personal factors with issues of evaluating test-based school ratings in the context of school and community factors. The article is focused on the latter topic, not the former.

Willson and Kellow say that measures such as percentages of students qualified for free or reduced-price lunch would be "more appropriate" than per-capita income for the purposes of the article—an opinion they don't explain or defend. The article shows that test scores are closely associated with incomes at all income levels found in the communities studied. Its results suggest that community income, as distinct from family income, may have substantial influence. A study of individuals would be needed to resolve the influences.

When typical incomes were well beyond poverty levels, as in many communities studied for the article, the percentage of high-school students qualified for free or reduced-price lunch became a poor proxy for income. A model using income directly was statistically much more effective. Willson and Kellow argue that this would be less of a problem for elementary-school students, but they neither cite nor present evidence. In a study focused on lower-income Florida communities Tschinkel, 1999, did find an association between test scores and "supported lunch" that was about as strong as the article finds between test scores and income for Massachusetts communities near Boston.

Willson and Kellow review MCAS tests and AREA/APS *Standards* (Committee, 1985, 1999), but the article does not invoke those or any other institutional standards. It is not focused on testing at a personal level, as standards are. Instead it is concerned with school ratings based on test scores. However, in this context Willson and Kellow are surely aware that jurisdictions such as California and Chicago ignored some of the AREA/APS standards in using commercial achievement tests for promotion and graduation tests. Other jurisdictions such as Texas and Massachusetts claim they comply with those standards but have legalistic interpretations.

Scattterplot data from Texas that Willson and Kellow show in their first figure could not be compared with data from Massachusetts in the article, because Willson and Kellow did not provide data, cite a source from which to obtain data, or translate their "economic disadvantage" index to income. They do not say whether their index reflects household or community income, nor do they evaluate the accuracy of the index as an income proxy for communities with typical incomes well above poverty.

There is a further, critical problem in trying to compare results from the Texas "accountability" testing program with those from Massachusetts. As many experiences show, teachers, students and parents adapt defensively to testing

programs—the higher the stakes, in general, the stronger the defenses. Education agencies also respond in a variety of ways to public reactions. Massachusetts patterns from 1998-1999 might be compared with Texas patterns from 1985-1986, when Texas testing began, but Massachusetts tests from 1998-1999 more closely resemble contemporaneous Texas tests than the Texas tests from 1985-1986.

Texas started "accountability" testing in 1985 and is now more than ten years into a second generation of tests (TAAS). It has been enforcing a graduation test requirement and maintaining its Accountability Rating System for more than eight years. By the late 1990s, there were widespread reports of weeks spent on test cramming, of "TAAS rallies," of heavy school spending on test prep materials and consultants, and of scandals over falsifying reports (see McNeil, 2000, for examples). Some observers such as Haney, 2000, suspect dropout rates have been redefined to conceal problems.

Massachusetts started "accountability" testing in 1998. Its graduation requirement will not directly affect any student until 2003. Massachusetts has some communities with a history of aggressive testing, such as Worcester; but before the school year ending in 2000 most schools made light to moderate responses to MCAS tests, and the Board of Education discounted most concerns about higher dropout rates.

Baseline data for "Significance of Test-based Ratings" were from 1998 and 1999, the quietest years of the MCAS program. Differences between those years were used to estimate variability, and the 1999 scores were used for most of the effects models. These data reflect conditions of Massachusetts schools before most strong responses. Scores from 2001, unavailable when the study was conducted, clearly show effects of strong responses, which will probably grow. Under state threats to take over or close low-scoring schools, there have already been heavy efforts to increase scores in some schools, involving test cramming that would be familiar to Texans; and there have been widely reported score increases.

In a short paragraph after their Texas scatterplot data, Willson and Kellow again object to correlating income with test scores, claiming income to be "uninterpretable." One is reminded of Tevye from *Fiddler:* "Impossible! Impossible!" Of course, the interpretation is entirely possible. Grissmer, et al., 2000, do it, the Nader organization does it, the Educational Testing Service does it, and if they try a bit harder, Willson and Kellow can do it too.

Willson and Kellow advocate teaching and learning metrics sometimes called the "value-added model" (e.g., by McLean, et al., 1998). Sanders has popularized a variant of this approach, and he supports it commercially (Sanders & Horn, 1995). The stability and significance of ratings based on such methods have recently been questioned by Kane & Staiger, 2001b, who also estimate contributions to score volatility from several sources.

Consistent with the article's observation (not "error") under Figure 2-10, Willson and Kellow also find year-to-year score changes exhibiting low significance. The Massachusetts Department of Education uses a score-change metric (Mass. DoE, 1999) slightly more robust than year-to-year changes. Kane & Staiger, 2001a, propose filters applied to several years of scores for better metrics. Willson and Kellow might compare score volatility estimates, by sources, that were obtained by Kane & Staiger, 2001b, from North Carolina elementary school testing against the within-year and between-year score variations from Texas testing.

Willson and Kellow complain that the article does not explore the content of MCAS tests. As previously stated, the article's focus is on significance of test scores for constructing school ratings, not on internal properties of the tests. However, the article provided Willson and Kellow with references to the full MCAS test content (Mass. DoE, 2000a, for 2000) and technical manual (Mass. DoE, 2000b, for 1999), which Massachusetts publishes on the Internet. They have had unlimited access to these documents for any reviews they find "appropriate."

In their Table 3, Willson and Kellow present a correlation matrix of score changes from Texas elementary schools plus social and school factors. Again it was not possible to compare Texas data with Massachusetts data from the article. The article associates factors with scores, while Willson and Kellow associate them with score changes for cohorts of identified students. There are also at least two major problems with results that Willson and Kellow present. First, they have score changes with high volatility. Obtaining a robust pattern would require multiple years, perhaps with filters such as Kane & Staiger, 2001a, propose. Second, they have social and school factors with substantial correlations, but they offer no multifactor model and no stepwise or combinatorial analysis of variance.

Willson's and Kellow's title for their critique echoes a shopworn slogan of the education hustlers: that school-based standard test scores are sending us a "message." Other than a bundle of sticks, what might that message be? A question rarely asked about "accountability" programs is whether their tests measure anything useful. What they measure is whether a test-taker, in a constrained situation, can interpret isolated fragments of information, solve small, arbitrary puzzles, recall miscellaneous items, or write in a simplistic style. Otherwise such situations may rarely be encountered—except perhaps with crossword puzzles or quiz shows.

Life's challenges are hardly ever so neatly packaged as the questions on a school-based standard test. They are often far more difficult: grasping people's real wants and needs, seeing advantages where others see limitations, organizing experience to make sense of it, understanding one's own blind spots, persisting against adversity, motivating people and guiding them. In most circumstances other than getting certificates that depend on results from those tests, how

would the results be of practical use? Attempts to measure education effectiveness using the current generations of state "accountability" tests may be mansions built on sand.

Associations of incomes with "aptitude" test scores have been recognized in the U.S. for more than 80 years. There are related studies about effects of poverty on cognitive development, such as Smith, et al., 1997, but the underlying behaviors at higher incomes are not understood much better now than they were in the 1920s. Flynn, 1984, showed that average IQ scores have been rising dramatically over time, suggesting that the underlying behaviors involve training or experience. Recently Dickens & Flynn, 2001, proposed an interpretive model, but so far little investigation of it has been reported. State "accountability" tests may be heavily weighted for the same cognitive skills that are sampled by "aptitude" tests, leading to associations with income like those that "Significance of Test-based Ratings" finds.

As this article shows, Massachusetts school ratings, based solely on MCAS test scores, are not precise educational measures of high significance. Score variations are large, and scores appear to reflect heavily the social advantages or disadvantages that students bring to schools from their backgrounds, not necessarily the effectiveness of the schools themselves.

References

Bolon, C. (2001). Significance of test-based ratings for metropolitan Boston schools. *Education Policy Analysis Archives 9*(42). October 16, 2001, available at http://epaa.asu.edu/epaa/v9n42.

Committee to Develop Standards for Educational and Psychological Testing, Novick, M. R., Chair (1985, revised 1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association and American Psychological Association.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: the IQ paradox resolved. *Psychological Review 108*(2), pp. 346-369.

Flynn, J. R. (1984). The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin 95*, pp. 29-51.

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving Student Achievement: What NAEP State Test Scores Tell Us*. Santa Monica, CA: RAND Corp.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives 8*(41) at http://epaa.asu.edu/epaa/v8n41.

Kane, T. J., & Staiger, D. O. (2001a). *Improving School Accountability Measures*. Cambridge, MA: National Bureau of Economic Research, Working Paper 8156.

Kane, T. J., & Staiger, D. O. (2001b). *Volatility in School Test Scores*. Washington, DC: Brookings Institution.

Massachusetts Department of Education (1999). *School and District Accountability System*.

Massachusetts Department of Education (2000a). *Release of Spring 2000 Test Items*. Available at http://www.doe.mass.edu/mcas/2000/release/.

Massachusetts Department of Education (2000b). *MCAS 1999 Technical Report*. Available at http://www.doe.mass.edu/mcas/2000/news/pdf/99techrep.pdf.

McLean, J. E., Snyder, S. W., & Lawrence, F. R. (1998). A school accountability model. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, Louisiana, November 4-6, 1998.

McNeil, L. M. (2000). *Contradictions of School Reform*, New York, NY: Routledge.

Milton, S (1999, February 25). Town lawyers endorse funding of RAGE lawsuit. *Cape Cod Times*, Barnstable, MA.

Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed. *Education Policy Analysis Archives 3*(6) at http://epaa.asu.edu/epaa/v3n6.html.

Smith, J. R., Brooks-Gunn, J., & Klebanov, P. K. (1997). Consequences of living in poverty for young children's cognitive and verbal ability and early school development. In G. J. Duncan & J. Brooks-Gunn (Eds.), *Consequences of Growing Up Poor* (pp. 132-167). New York, NY: Russell Sage Foundation.

Texas education funding challenged in new lawsuit (1998, May 8). *Daily Tax Report*. Washington, DC: Bureau of National Affairs.

Tschinkel, W. R. (1999, April 25). Poverty, not bad schools, hinders learning. *Miami Herald*.

About the Author

Craig Bolon
Planwright Systems Corp
P.O. Box 370
Brookline, MA
02446-0003 USA
Email: cbolon@planwright.com
Phone: 617-277-4197

Craig Bolon is President of Planwright Systems Corp., a software development firm located in Brookline, Massachusetts, USA. After several years in high energy physics research and then in biomedical instrument development at M.I.T., he has been an industrial software developer for the past twenty years. He is author of the textbook *Mastering C* (Sybex, 1986) and of several technical publications and patents. He is an elected Town Meeting Member and has served as member and Chair of the Finance Committee in Brookline, Massachusetts.

151

Women in Managerial Positions in Greek Education: Evidence of Inequality

Anastasia Athanassoula-Reppa
Technical and Vocational Teacher Training Institute of Greece

Manolis Koutouzis
Hellenic Open University
Greece

Abstract

This article deals with the under-representation of women in managerial positions in Greece. While substantial progress has been made in terms of the legal framework that ensures equal rights to both men and women in the country, evidence shows that there are barriers that inhibit women from pursuing and taking such positions, resulting to covert discrimination. This occurs despite the dominance of women in Greek education. We regard that kind of discrimination as a democratic deficit; it contradicts the notion of "democratic citizenship." Although we do not advocate a quota system, we stand for implementation of basic democratic principles, which could prevent such discrimination.

Introduction

The dawn of the 21st century has brought once again the issue of citizenship to the front of socio-political arguments. The failure, within the context of globalization, of both social-democratic "statism" and Thatcherite free-market economies to resolve burning issues such as unemployment and social exclusion, has led to the reconsideration of the "Civil Society" and the post-modern "Citizen" (Cohen & Arato, 1992). Traditionally, citizenship has been considered to have two principal dimensions: the *civic,* concerned primarily with the fundamental freedoms of speech, thought and religion, and the *political,* concerned with participation in political developments and the right to vote and be elected (Marshall, 1995, Tilly, 1995).

However, the realization that crucial decisions about the development of postmodern globalized societies are made without the actual participation of the citizens and that large sectors of these societies are excluded from fundamental social rights has sensitized citizens to the issues of participation and exclusion. Thus, another dimension has been added to the concept of "citizenship," namely, the *social dimension.* It includes fundamental social rights such as

152

access to health, work, welfare and participation in decision-making mechanisms. These rights, when excersized on an equal basis by all members of the society, are now considered to be at the heart of democracy. Recent developments in Seattle, Gothenburg and Genoa illustrate the reaction caused when parts of the (globalized) society feel deprived of the opportunity to participate in decision-making processes.

Arguably, access to the rights that constitute the social dimension of citizenship has not been gained simultaneously by both men and women; indeed, it has been considered a "male-privilege" for many societies. For many years, liberal democracy had built the notion of "citizenship" using male stereotypes as a basis (James, 1996, Walby, 1994, Kavounidi, 1998). Issues such as occupational choice and access to various professions, participation in decision-making positions at work or in other aspects of social life, promotion criteria, are traditionally not only determined by personal preference and psychological motives, but also related to historical, sociopolitical, ideological and cultural mechanisms. According to researchers, these mechanisms have been gender biased, at least to a certain extent. (Kassimati, 1989, Eliou, 1993, Vassilou – Papageorgiou, 1995, Kaltsogia-Tournavitou, 1997). In other words, women have not had equal access to these rights. Undoubtedly, however, research and analysis of such phenomena of inequality and their origins pose substantial difficulties and certainly go beyond the scope of the research reported here.

In this article, we focus on the observed under-representation of women in decision making mechanisms of the Greek educational system, and more specifically in school management. We attempt to identify the reasons behind that phenomenon, which we consider a clear example of the limited development of the *social dimension of citizenship* in Greece. The data used for this research were provided by the Greek Ministry of Education and have been analyzed by the authors.

Women in the Greek Labor Market

Consider the general picture of Greek labor market and the position of women in it. According to 1996 statistics, the workforce in Greece is about 4.3 million persons, of which 2.6 million are men (60%) and 1.6 million are women (40%). The percentage of women in high-ranking positions, however, does not match the above overall distribution. Only one of the political-parties has a woman as a leader (the Greek Communist Party), while only 5.6% of the Members of Parliament (MPs) and 16% of the Euro-MPs are women. Women in governmental positions have never exceeded 12%. Furthermore, although 36% of the people working in the media are women, only 10% are in managerial positions. To put it plainly, women are not proportionally represented in high-ranking, prestigious positions. According to Damoulianou (1998), despite the fact that for more than 15 years there are more women than men studying in the Greek Universities, this predominance of women in higher education is not reflected in the labor market, where inequality is observable in quantitative as well as qualitative terms. The above examples are just quantitative evidence that inequality persists in the Greek society and institutions, as in other Western European countries.

According to Eurobarometer (1998), within the EU, the level of female participation in positions of "high responsibility" is considerably low. The reasons according to the same source include following:

- Lack of time, due to family responsibilities;
- The working environment is male-dominated and does not "trust" women;
- Women are not "ready to fight" for their careers;
- Women do not always possess the necessary psychological characteristics to cope with the pressures of such a male-dominated environment;
- Women are "not interested" in such positions.

While the above are said to be typical of all the EU countries, they are definitely valid for Greece, as relevant studies have shown. In the following paragraphs, we hope to demonstrate that the profession of education evidences clearly the validity of the above argument. What is noteworthy is that the domain of education is **not** male dominated in Greece, as it is shown in the following paragraphs.

Women in Education

It has been suggested by numerous researchers and by statistical data, that in most developed countries women are over-represented in pre-primary and primary education as well as in general secondary education as opposed to technical and vocational secondary education. (Wilson, 1997). In Greece, the same pattern is evidenced as shown in Table 1.

Table 1
Women in Greek Education (1997-1998)

	Number of Teachers	Number of Women Teachers	Percentage of Women Teachers

153

Nursing Schools	8,785	8,604	97.9
Primary Education	45,814	25,572	55.8
General Secondary	49,733	29,225	58.8
Technical-Vocational Secondary	19,069	8,083	42.4
TOTAL	**123,401**	**71,484**	**57.9**

Source: Ministry of Education, Annual Statistics, 1998

Numerous research projects and authors have tried to explain the phenomenon of high representation of women in the teaching profession (both in primary and secondary education), not only in Greece but in Western societies in general (Vassilou – Papageorgiou, 1992, Bucher & Saran, 1995, Cowan & Koutouzis, 1997, Dimitropoulos, 1997, Neave, 1998). We shall not repeat the arguments here. Briefly however, teaching (especially in primary and lower secondary education) has been considered to be the continuation of childcare and child-rearing, which in turn has been associated with women in the above societies. Given these facts, however, one would expect a strong representation of women in managerial positions in the Greek educational system.

Women in Managerial Positions—The Case of Greece

There are four kinds of managerial positions in the Greek educational system: **Heads of Schools, Heads of Regional Educational Office** (Local Educational Authorities), **School Advisors,** and finally **Heads of Greek Educational Offices Abroad.** What has to be noted however, is that the responsibilities of all the above are limited compared to related positions in other educational systems. The highly centralized nature of the system is mainly responsible for the limited authority of the above positions.

For all schools, the center has: defined the content of the national curriculum; recommended appropriate teaching methods; published textbooks; allocated funding; legislated for participation by various stakeholders in schooling; determined student examinations and has taken full responsibility for the organization of schools, including all aspects of staffing (OECD, 2001 p. 79).

Traditionally, within the highly centralized and bureaucratized Greek educational system, Heads of Schools have been administrators expected to follow and implement decisions made at the central level, i.e., the Ministry of Education. The same could be argued for the Heads of Regional Education Offices (HREOs). They are the "link" between central government and local schools, and they coordinate the schools in their area of responsibility. They are responsible for allocating staff to the schools of the region for which they are responsible. However, the Ministry of Education allocates the staff to the region. In essence they do not decide the number of teachers; they merely administer the decision made at the central level.

School Advisors (SA) have a slightly different role. They are experienced subject specialists, often holding post-graduate degrees, and they assist teachers by offering advice and disseminating good practices. The area of responsibility of each School Advisor depends on the number of teachers teaching the specific subject in each region. For instance there is only one Advisor for Art Education for the whole of the country, but several for Mathematics (34) and Language & Literature (98). Their role is also centrally determined.

Despite the limitations posed by the educational system, all three positions enjoy a certain degree of prestige and status in Greek society, with the HREO and the SAs placed higher in the "ranking." This can be explained by the crucial role of the schools in the early stages of the formation of Greek society, and also by the participation of HREO and SA in various Assessment and Selection Committees. As discussed below, these positions, irrespective of their status in the society, are key to the successful operation of the system and thus, important for the Ministry (and the Minister) of Education.

Perhaps the most prestigious, but definitely the most attractive and well-paid positions, are the Heads of Greek Educational Offices Abroad (HGEOA). Their role is *advisory, managerial,* and *controlling.* In essence, they represent the Greek Ministry of Education in their area of responsibility. Occasionally, HGEOA have to negotiate with authorities of the host country on issues of organization and administration of the Greek schools, while they also play an important role in the social life of the Greek Diaspora. We could argue, therefore, that their role is also political. There are 26 such positions around the world: 13 in Western Europe and 13 in other continents (USA, Canada, Argentina, Egypt, South Africa, Ukraine, Turkey and 6 in Australia). HGEOA enjoy greater autonomy than their counterparts in Greece, as it is difficult for the Ministry to interfere in everyday aspects of Greek education in these areas. Their selection, however, is centrally administered by the Ministry of Education.

According to Table 2, 41% of Primary Heads are women, although almost 56% of primary teachers are women.

Table 2
Percentage of Women in School Management in European Countries

Country	Primary Education		Secondary Education	
	Teachers (%)	Heads (%)	Teachers (%)	Heads (%)
England and Wales	81	49	49	26
France	79	64	56	30
Greece	56	41	59	36
Hungary	85	33	97	30
Ireland	78	46	54	29
Italy	93	46	63	30
Netherlands	76	13	33	7
Norway	74	40	39	22
Spain	74	47	50	20

(Source: Wilson, 1997)

What is interesting in the above table is that Greece stands among the countries with the highest representation of women in School Management positions, indicating that the reasons for the observed under-representation are not country-specific. Rather they come as a result of the reasons stated above by the Eurobarometer study.

The relatively high percentage of women appointed as Heads of schools in Greece does not in any case mean that equality has been achieved, or that women have equal access to such positions. It just shows that the phenomenon of under-representation of women in managerial positions is not unique to Greece.

If we now turn to the more prestigious and, arguably, more influential positions of Heads of Regional Education Offices and School Counselors, the phenomenon of inequality and under-representation is clearly demonstrated. According to Tables 3 and 4, during the last selection process in 1998, only 11 women primary teachers out of 443 candidates expressed an interest in becoming Heads of Regional Office. In secondary education, the numbers were 13 out of 466.

Table 3
Heads of Regional Education Offices (Primary Education)

	Total	Women	% of Women
Candidates	433	11	2.48
Selected	199	6	3.01

Source: Ministry of Education, Annual Statistics, 1998

Table 4
Heads of Regional Education Offices (Secondary Education)

	Total	Women	% of Women
Candidates	466	13	2.78
Selected	191	13	6.81

Source: Ministry of Education, Annual Statistics, 1998

What is interesting to note is that all 13 women selected in secondary education were Greek language and literature

155

teachers.

Moving to the School Advisors (Table 5), we notice that in Primary Education there were only 72 out of 512 women candidates, and in Secondary Education 148 out of 625.

Table 5
School advisors in Greece

	Candidates			Selected		
	Total	Women	%	Total	Women	%
Pre-Primary Education	104	104	100	49	49	100
Primary Education	512	72	14.1	301	32	10.6
Special Education	57	8	14	16	0	0
Secondary Education	625	148	23.7	254	47	18.5
Total	1,298	332	25.6	620	128	20.6

Source: Ministry of Education, Annual Statistics, 1998

Finally, during the last selection process, for the position of Head of Greek Educational Offices Abroad, there were 199 candidates, 140 men and 59 women as we see in Table 6. Only 5 women were selected.

Table 6
Heads of Greek Educational Offices Abroad

	Total	Women	% of Women
Candidates	199	59	29.6
Selected	26	5	19.2

If we now compare the above figures with the percentage of women teachers in Greece, we can easily reach the conclusion that female participation in managerial or other "crucial" positions in the Greek educational system is not as high as expected and does not reflect the composition of the teaching profession in the country. We see four interrelated reasons for this under-representation. There are three levels of overt or covert discrimination that result in the unequal representation of women in such high-status and highly responsible positions. Below we attempt to identify these levels of discrimination and the main reasons for them.

Reasons for the low participation of women in managerial positions

Personal—psychological barriers

Research has shown that lack of interest on behalf of women in managerial or other highly responsibile positions can be explained by the stress caused by role-conflict (Al Khalifa, 1992, Thompson, 1992). A woman teacher does not separate her working life from her "personal" life in the same way a man does. Worrying that such a position and responsibility may absorb time dedicated to her family, she is reluctant to apply for it. It is felt that having "two jobs" places a significant burden on women's shoulders no matter how helpful their partners are (Singleton, 1993).

In the Greek context, part of the responsibilities of some School Advisors is to travel and advise teachers from different areas, even prefectures. If a woman takes on such a responsibility, given the family structures in Greece, she is definitely aware that such a decision would probably cause "disorder" within her family. Thus, she is intrinsically demotivated, and prefers to continue her career as an ordinary teacher. Research has shown that women feel more satisfied in the teaching profession than men do, as they feel that there is no incompatibility between their personal and working life (Dimitropoulos, 1997).

Psychological barriers are not expressed only in a lack of interest for managerial positions. It has also been argued that women feel they should also adopt "male behavior" in order to become accepted and appreciated in such positions (Shakeshaft, 1987). Such an argument, however, is not valid, as it has been heavily disputed by research evidence. According to Shakeshaft, (1987) and Robetrson, (1996), in schools headed by women academic achievement and morale is higher, there is less violence, and generally fewer discipline problems. Also, De Lyon and Migniuolo (1989), confirming the above argument, suggest that it is women's rather different approach to educational

156

management that succeeds. Moreover, current discussions about management and leadership in schools bring to the surface the effectiveness of more democratic, flexible and participatory models of leadership (Koutouzis, 1999). Such models do not require the dominance of a male Headmaster but rather the skill to bring together views and opinions. "The very nature of management, dealing as it does with areas of uncertainty, negotiation and policy making, draws on feminine qualities of intuition, aesthetic considerations, dependence on colleagues and so on" (Singleton, 1993, p.175). This is not to say that all women in relevant positions adopt such a leadership style. It indicates, however that the male stereotype is not the only way to efficiency and effectiveness.

Institutional barriers

By the term "institutional barriers," we mean all barriers related to the educational system, its structure and the way it is organized and managed. As mentioned above, all three positions are considered crucial in terms of political operation of the Greek educational system, given its highly centralized and bureaucratized nature. We could argue, therefore, that authorities, irrespective of their political stand and ideology, seek to manipulate the crucial positions of the system by defining the selection criteria and controlling the internal structure of the system. The final aim is to be able to promote and realize educational policies determined at the center, i.e., the Ministry of Education. The fact that in every governmental change there is also an imposed change of persons in the above- mentioned positions can only confirm the argument of political manipulation of the system and its key posts.

Following the above argument it is safe to assert that such manipulation can be associated with gender issues. Let us be more specific. The selection criteria for all three key positions under discussion can be divided into two main categories: a) objective criteria, and b) subjective criteria.

In the first category, all academic or other qualifications, which can be proved by relevant degrees, certificates and the like are included. Professional and other managerial experience is also included in it. In the second category, personal skills and qualities are included. Ability to lead and manage, general social activity, participation in local clubs... are among the expected qualities, assessed by the –centrally appointed – Selection Committee. The assessment of the Committee is of utmost importance in cases where the qualifications presented are about equal.

The composition of the Committee has always been male dominated. Given the persisting stereotypes in Greek society (see below), we would expect that in cases of equal qualifications, male candidates are preferred.

Social—cultural barriers

It is very crucial for our argument to appreciate that in Greece gender equality has been introduced into legislation fairly recently. The Constitution of 1975 established legal equality between men and women. According to it, men and women in Greece enjoy the same rights and responsibilities in all aspects of social life (education, work, healthcare, etc.). However, it was only in 1983 that institutional and legal "barriers" were truly removed, establishing gender equality. For nearly a decade, these barriers, due to lack of subsequent relevant legislation, prohibited the realization of the constitutional right of equality (Kaltsogia-Tournavitou, 1997). "Jobs whilst not legally labeled 'for men' or 'for women' are still viewed by many people as just that"(Singleton, 1993, p.165).

It would be safe to argue, therefore, that the process of reaching gender equality in Greece started less than twenty years ago. The results of the process and, more importantly, subsequent changes of attitudes and cultural norms, are not immediate; and evidence of inequality - hidden rather than overt – can be observed in many aspects of Greek social life even today. We observe in Greece, the phenomenon described elsewhere: substantial equality can not be achieved as long as "hidden" discrimination and preferences are reproduced. It is not enough to declared equality if you "do not feel very comfortable facing a woman in a authority position" (Al Khalifa, 1992).

In the area of educational management attitudes and perceptions, follow the patterns described above. "Somehow people assume that men possess the necessary qualities to do the job and this only changes when they demonstrate otherwise, but with women, we have to prove over and over again that we can do the job before our abilities are recognized" (Singleton, 1993, p.171) . Quite simply, educational management is considered a "male" job, not only by society in general but also by teachers and even pupils. Research evidence confirms that pupils hold a preconception that effectiveness of the school is increased by having a male as Headteacher, who tolerates less "mucking about" (Stanworth, 1984).

Conclusion and Implications

Despite considerable progress made in various aspects of Greek political and social life in general, there is clear evidence of female under-representation in managerial positions of the Greek educational system. The reasons for this are traced not solely to the socio-cultural barriers that persist in Greek society. Personal as well as institutional barriers complete a picture of covert discrimination. The fact that the teaching profession is "dominated" by women has not resulted to equal representation in positions of relatively higher status and responsibility. Such discrimination, as we stated in the beginning, is not just evidence of male-female inequality and unfair treatment. It goes far beyond that, and it is rather a clear sign of violation of democratic attitudes and practices in a democratic country. It

demonstrates the exclusion (overt or covered), of a significant part of the society from certain positions and the weakening of the *social* dimension of citizenship. The fact that the same phenomena are observed elsewhere in the western world does not weaken our argument. On the contrary, it confirms the existence of a democratic deficit in the western world where large sectors of society are excluded from decision-making positions and mechanisms.

In an era that calls for greater participation of all parts of society in social and political developments, in an era that has demonstrated that observed democratic deficits create tensions, the covert exclusion of majorities from decision-making mechanisms and positions is clearly not acceptable. While other authors propose the use of quotas to improve the situation, we strongly advocate respect for fundamental democratic principles.

References

Al Khalifa, E. (1992). Management by halves: women teachers and school management. In Benett, N., Crawford, M. & Riches, C. (eds). (1992). *Managing Change in Education: Individual and Organizational Perspectives*. London: Paul Chapman Publishing.

Andreou, A. & Papakonstantinou G., (1994), *Εξουσία και Οργάνωση– Διοίκηση του Εκπαιδε υτικού Συστήματος*. Αθήνα: Νέα Σύνορα – Α. Α. Λιβάνη,

Benett, N., Crawford, M. & Riches, C., (eds), (1992), *Managing Change in Education: Individual and Organizational Perspectives*. London: Paul Chapman Publishing,.

Bucher H. & Saran, R. (1995). *Managing Teachers as Professionals in Schools*. London: Kogan Page.

Cohen, J.L. &Arato A. (1994). *Civil Society and Political Theory*. Cambridge: MIT Press.

Coleman, M., (1994). Women in Educational Management. In Bush T. & West-Burnham J. (1994). *Principles of Educational Management*. London: Longman.

Cowan B.J. & Koutouzis E. (1997). Mismanagement Ambiguity and Delusion: Training Primary Teachers in Greece. *Mediterranean Journal of Educational Studies, 2* (2).

Damoulianou, H., (1998) " Τα Συνδικάτα 'ξέχασαν' τις γυναίκες" Καθημερινή,22/2/1998.

De Lyon, H. & Migniulo, F. (1989). *Women Teachers*.Milton Keynes: Open University Press.

Dimitropoulos E., (1997), Οι εκπαιδευτικ&omi cron;ί και το επάγγελμά τους. Αθήνα: Γρηγόρης.

Eurobarometer, May, 1998

Eliou, M., (1993), Εκπαιδευτικ^ 2; και Κοινωνική Δυναμική. Αθήνα: Πορεία.

Gray, H.L. (1989). Gender considerations in school management: masculine and feminine leadership styles. In Riches, C. & Morgan, C. (eds). (1989). *Human Resource Management in Education*. Milton Keynes: Open University Press.

James, S., (1996), "Καλοί και καλύτεροι πολίτες: Η ιδιότητα του πολίτη και η γυναικεία ανεξαρτησία", Δίνη, 8.

Kavounidi, T., (1998), "Κοινωνικός Αποκλεισμός, Ιδιότητα του πολίτη και φύλο" στα Πρακτικά Συνεδρίου Κοινωνικές Ανισότητες και Κοινωνικός Αποκλεισμός, Αθήνα: Ίδρυμα Σάκη Καράγιωργα.

Kaltsogia – Tournavitou, N., (1997) *Υποαντιπροσα 4;πευση Γυναικών και Δημοκρατία – Η συνταγματικό& tau;ητα των θετικών μέτρων για την ενδυνάμωση του ρόλου των γυναικών στη δημόσια ζωή*. Αθήνα: Σάκκουλας.

Kassimati, K., (1989) Επιλογή Επαγγέλματο ς: Πραγματικότητ&al pha; ή Μύθος; Αθήνα: ΕΚΚΕ.

Kontogiannopoulou – Polydorides G., Zambeta, E. (1997). Women in Educational Management in Greece. In Wilson M., (ed). (1997). *Women in Educational Management: A European Perspective*. London: Paul Chapman Publishing.

Koutouzis, M., (1999), Βασικές Αρχές Management. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.

Marshall, T.H., (1995), Ιδιότητα του Πολίτη και Κοινωνική Τάξη. Αθήνα: Gutenberg,.

Ministry Of Education (Greece), Annual Statistics, 1998.

Neave G., (1998), *Οι Εκπαιδευτικ&omi cron;ί. Προοπτικές για το εκπαιδευτικα 2; επάγγελμα στην Ευρώπη*. Αθήνα: Έκφραση.

OECD, (2001), *What Works in Innovation in Education- New School Management Approaches*. Paris:OECD.

Preedy, M., (ed) (1993), *Managing the Effective School*. London: Open University - Paul Chapman Publishing.

Robertson, H.,L., (1995), "Η εξέλιξη των εκπαιδευτικα 4;ν και η δίκαιη αντιμετώπιση των δύο φύλων", στο Hargreaves, A., Fullan, M., (επιμ.), (1995), Η *Εξέλιξη των Εκπαιδευτικα 4;ν*. Πατάκης, Αθήνα.

Shakeshaft, C.. (1987). *Women in Educational Administration*. Newbury Park, CA: SAGE.

Shakeshaft, C.. (1993). Women in educational management in the United States. In Ouston, J. (ed.). *Women in Education Management*. Boston: Longman.

Singleton, C.. (1993). Women Deputy Headteachers in Educational Management. In Preedy, M., (ed) (1993). *Managing the Effective School*. Open University, London: Paul Chapman Publishing.

Stanworth, M.. (1984). *Gender and Schooling*. London: Hutchinson &Co.

Thompson, M. (1992). Appraisal and Equal Opportunities. In Benett, N., Crawford, M. & Riches, C. (eds). (1992). *Managing Change in Education: Individual and Organizational Perspectives*. London: Paul Chapman Publishing.

Tilly, C. (1995). Citizenship, identity and social history. *International Review of Social History, 40*, Supplement 3.

Vasilou – Papageorgiou V., (1992), "Εκπαίδευσ η και φύλο: Η διάσταση του φύλου στο εκπαιδευτικα 2; προσωπικό της πρωτοβάθμιας εκπαίδευσης.& rdquo; Στο Καζαμίας Α. & Κασσωτάκης Μ., (1995) Ελληνική Εκπαίδευση: Προοπτικές Ανασυγκρότη&sigma ;ης και Εκσυγχρονισ& mu;ού, Σείριος, Αθήνα.

Walby, S. (1994). Is citizenship gendered? *Sociology, 28*(2).

Wilson M. (ed). (1997). *Women in Educational Management: A European Perspective*. London: Paul Chapman Publishing.

About the Authors

Anastasia Athanssoula Reppa
Technical & Vocational Teacher Training Institute of Greece
Athens, Greece

Anastasia Athanssoula Reppaholds a Ph.D. from Panteion University, Athens. She teaches Educational Management and Administration in Technical and Vocational Teacher Training Institute of Greece. She has also been an adjunct lecturer at the University of Athens and a tutor at the Hellenic Open University. She has published a number of books and papers related to educational Management and administration. Her latest area of interest is gender issues in educational administration.

Manolis Koutouzis
National Center for Educational Research
Athens, Greece

Email: mkou@kee.gr

Manolis Koutouzis holds a Masters and Ph.D. from the University of Reading, England. He is a Research Associate at the Center for Educational Research, Athens and a Tutor at the Hellenic Open University. He has worked as a country expert for the OECD project (2001): "What Works in Innovation in Education: New School Management Approaches". His areas of interest include educational management, educational planning and policy-making.

Departamento de Investigación Educativa-
DIE/CINVESTAV
rkent@gemtel.com.mx   kentr@data.net.mx

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Universidad Federal de Rio Grande do Sul-UFRGS
luccmb@orion.ufrgs.br

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

161

How Schools Matter:
The Link Between Teacher Classroom Practices
and Student Academic Performance

Harold Wenglinsky
Educational Testing Service

Abstract
Quantitative studies of school effects have generally supported the notion that the problems of U.S. education lie outside of the school. Yet such studies neglect the primary venue through which students learn, the classroom. The current study explores the link between classroom practices and student academic performance by applying multilevel modeling to the 1996 National Assessment of Educational Progress in mathematics. The study finds that the effects of classroom practices, when added to those of other teacher characteristics, are comparable in size to those of student background, suggesting that teachers can contribute as much to student learning as the students themselves.

Introduction

Much of the discussion in educational reform hinges on the question of whether schools matter. Over the past two decades, policymakers have called for improvements in the academic performance of U.S. students. Many educational reformers, particularly those associated with the standards movement, hold that the key to improving student performance lies in improving the schools. If academic standards are rigorous, curriculum and assessments are aligned to those standards, and teachers possess the skills to teach at the level the standards demand, student performance will improve. However, this perspective is to some extent at odds with another that has emerged from the discussion about school improvement, namely that it is students rather than schools that make the difference. Hence, a *New York* Times story on how to improve the academic performance of low-income students can include the headline: "What No School Can Do (Traub, 2000)." Or, as Laurence Steinberg puts it in *Beyond the Classroom: Why School Reform has Failed and What Parents Need to* Do, "neither the source of our achievement problem, nor the mechanism through which we can best address it, is to be found by examining or altering schools (Steinberg, 1996, p. 60)." In this view it is the social backgrounds of students that play the key role in their ability to learn, and only by moving outside of the educational system and attacking the pervasive economic inequalities that exist in the U.S. can student performance be improved.

Quantitative research on whether schools matter has generally supported the notion that the problems of U.S. education lie outside of the schools. Some research finds that when the social backgrounds of students are taken into account, school characteristics do not seem to influence student outcomes, suggesting that schools do not serve as avenues for upward mobility, but instead reinforce existing social and economic inequalities (Coleman et al., 1966; Jencks et al., 1972). Other researchers contend that school characteristics can have a greater effect on student

162

outcomes than would be expected based upon student background (Lee, Bryk and Smith, 1993). But while the research in support of this contention does find significant effects for school characteristics, the magnitudes of these effects tend to be modest, far overshadowed by the effects of student background characteristics.(Note 1)

A possible reason for the lack of large school effects in quantitative research is the failure of such research to capitalize on an insight from qualitative research: the central importance of the classroom practices of teachers. As far back as Willard Waller (1932), qualitative researchers have noted that the interaction which occurs between teachers and students in the classroom is greater than the sum of its parts. Students can leave the classroom with their knowledge and attitudes dramatically altered from what they were before they entered. Quantitative research neglects this dimension of schooling by treating it as a "black box," not worthy of study (Mehan, 1993). Often teaching is not studied at all, and, when it is, only the characteristics of teachers that are easily measured but far removed from the classroom (such as their level of educational attainment) are included.

The current study seeks to fill this gap in the literature by using quantitative methods to study the link between student academic achievement and teacher classroom practices, as well as other aspects of teaching such as the professional development teachers receive in support of their classroom practices and the more traditional teacher background characteristics, referred to here as teacher inputs. Such a study is made possible by the availability of a large-scale nationally representative database, the National Assessment of Educational Progress (NAEP), which includes a comprehensive set of classroom practices along with student test scores and other characteristics of students and teachers. For this study, the 7,146 eighth graders who took the 1996 assessment in mathematics are studied along with their mathematics teachers. The statistical technique of multilevel structural equation modeling (MSEM) is employed to address the major methodological shortcomings of the quantitative literature, namely the failure to distinguish between school- and student-level effects, to measure relationships among independent variables, and to explicitly model measurement error. The study finds that classroom practices indeed have a marked effect on student achievement and that, in concert with the other aspects of teaching under study, this effect is at least as strong as that of student background. This finding documents the fact that schools indeed matter, due to the overwhelming influence of the classroom practices of their teachers.

Background

Much of the quantitative literature linking school characteristics to student outcomes focused on the impact of economic characteristics, or school resources. These studies are known as production functions. One of the earliest of these studies was the Equality of Educational Opportunity Study, commonly referred to as the Coleman Report (Coleman et al., 1966). This study applied Ordinary Least Squares (OLS) regression analysis to nationally representative samples of elementary and secondary school students to relate school resources such as per-pupil expenditures to student academic achievement and other outcomes. The study found that, on average, when student background was taken into account, school resources were not significantly associated with student outcomes. Nearly 400 additional production function studies have since been conducted. Meta-analyses tabulating the results of such studies between 1964 and 1994 reached divergent conclusions. Some concluded that these studies showed no consistent relationship between school resources and student achievement (Hanushek, 1997, 1996a, 1996b, 1989), while others concluded that the studies showed a consistent, albeit modest, positive relationship (Greenwald, Hedges, & Laine, 1996; Hedges & Greenwald, 1996; Hedges, Laine & Greenwald, 1994).(Note 2)

Another line of inquiry into the impact of schooling on students, focusing on the social and organizational characteristics of schools, also emerged from the Coleman Report. This body of research, known as effective schools research, sought to identify common characteristics of schools in which students performed above what would be expected based upon their backgrounds (Edmonds, 1979; Brookover et al., 1979; Austin & Garber, 1985). While the earliest of these studies tended to be small in scope, later studies using large-scale databases confirmed many of their basic findings (Lee, Bryk & Smith, 1993; Chubb & Moe, 1990). These studies found that such characteristics of schools as the leadership qualities of the principal, the disciplinary environment of the school and the size of the student body all had an effect on student outcomes. In comparison to student background, however, these effects appeared quite modest.

Much of the quantitative research which focused specifically on teaching conformed to a similar pattern, finding little relationship between teacher inputs and student achievement. The Coleman Report measured seven teacher characteristics: years of experience, educational attainment, scores on a vocabulary test, ethnicity, parents' educational attainment, whether the teacher grew up in the area in which he or she was teaching, and the teacher's attitude toward teaching middle class students. For most students, this study found these characteristics to explain less than 1% of the variation in student test scores. The findings of the meta-analyses of production function studies were just as mixed for teacher inputs as for other school resources. They found that less than one-third of the studies could document a link between student outcomes and teacher experience, less than one-quarter could do so for teacher salaries, and just one in ten could do so for educational attainment; from such mixed results, the meta-analyses came to divergent conclusions, some suggesting a positive relationship and some suggesting no relationship.

More recent research on teaching has confirmed the lack of a clear relationship between student outcomes and teacher inputs, but with two exceptions: the amount of coursework the teacher had pursued in the relevant subject area and the teacher's scores on basic skills tests. Two analyses of large-scale databases revealed that exposure teachers received to college-level courses in the subject they were teaching led to better student performance. Monk (1994) analyzed 2,829 high school students from the Longitudinal Study of American Youth. These students were tested in mathematics and

science in 10th, 11th and 12th grades, and filled out questionnaires on their background characteristics. Their mathematics and science teachers were also surveyed. The study related teacher characteristics to student test scores, taking into account students' earlier test scores, background characteristics and teacher inputs. The study found that the more college-level mathematics or science courses (or math or science pedagogy courses) teachers had taken, the better their students did on the mathematics and science assessments. The more traditional teacher inputs that had been measured in the earlier production function studies, such as teacher experience or educational attainment, proved unrelated to student achievement. Similar results were obtained in a study by Goldhaber and Brewer (1995). They analyzed data on 5,149 10th graders, 2,245 mathematics teachers and 638 schools drawn from the National Educational Longitudinal Study of 1988 (NELS:88). Of the various inputs studied, the only one found to make a difference was the proxy for college-level mathematics coursetaking, namely whether the teacher had majored in mathematics.

Another series of recent studies suggested that, in addition to the teacher's coursework in the relevant subject making a difference, so too did the teacher's proficiency in basic skills as measured by standardized tests. Ferguson (1991) analyzed data on nearly 900 Texas districts, representing 2.4 million students and 150,000 teachers. He related the district average of various teacher inputs to average student scores on a basic skills test, taking into account student background. All of the school variables taken together accounted for from 25% to 33% of the variation in average student test scores, and one input, teachers' scores on the Texas Examination of Current Administrators and Teachers, a basic skills test, accounted for the lion's share of this effect. Similar results were obtained by Ferguson and Ladd (1996) in their study of Alabama school districts. Another district-level analysis, this time of 145 North Carolina school districts (Strauss & Sawyer, 1986), found a relationship between average teacher scores on a licensure test, the National Teacher Examination, and student scores on two different assessments taken by high school juniors, taking into account other school and student characteristics. The Coleman data have even been reanalyzed, finding a link between teacher scores on a vocabulary test and student scores on tests in various subject areas (Ehrenberg & Brewer, 1995). That study aggregated data to the school level, analyzing samples of 969 elementary and 256 secondary schools. The study calculated a dependent variable, a "synthetic gain score," as the difference between mean student scores in the sixth and third grades for elementary school students and in the twelfth and ninth grades for high school students. The study related teachers' educational attainment; experience and scores on a vocabulary test to synthetic gain scores and found only the latter to be consistently related to student performance.

Although large-scale quantitative research studied those aspects of teaching that are easily measurable, such aspects tend to be far removed from what actually occurs in the classroom. To study teacher classroom practices and the kinds of training and support pertinent to these practices which teachers receive, it is necessary to draw primarily on the findings of qualitative research.

The qualitative literature on effective teaching emphasizes the importance of high-order thinking skills (McLaughlin & Talbert, 1993). Teaching higher-order thinking skills involves not so much conveying information as conveying understanding. Students learn concepts and then attempt to apply them to various problems, or they solve problems and then learn the concepts that underlie the solutions. These skills tend to be conveyed in one of two ways: through applying concepts to problems (applications) or by providing examples or concrete versions of the concept (simulations). In either case, students learn to understand the concept by putting it in another context. In the case of an application, this might mean solving a unique problem with which the student is unfamiliar. In the case of a simulation this might mean examining a physical representation of a theorem from geometry or engaging in a laboratory exercise that exemplifies a law from chemistry. While both lower-order and higher-order thinking skills undoubtedly have a role to play in any classroom, much of the qualitative research asserts that the students of teachers who can convey higher-order thinking skills as well as lower-order thinking skills outperform students whose teachers are only capable of conveying lower-order thinking skills (see also Phelan 1989; Langer & Applebee, 1987).

The qualitative research also emphasizes three additional classroom practices: individualization, collaboration and authentic assessment. Individualization means that teachers instruct each student by drawing upon the knowledge and experience that that particular student already possesses. Collaborative learning means that teachers allow students to work together in groups. Finally, authentic assessment means that assessment occurs as an artifact of learning activities. This can be accomplished, for instance, through individual and group projects that occur on an on-going basis rather than at a single point in time (McLaughlin & Talbert, 1993; Graves & Sunstein, 1992; Golub, 1988).(Note 3)

The qualitative research suggests that this set of classroom practices can produce qualitative improvements in the academic performance of all students, regardless of their backgrounds. The focus on higher-order thinking skills is not only appropriate for advanced students; even those in need of more basic skills can benefit from understanding the conceptual basis of these skills. And individualization of instruction does not simply mean using special techniques for low performing students; techniques developed to address the problems of low-performing students can often help high-performing students as well. Regardless of the level of preparation students bring into the classroom, the qualitative research asserts, decisions that teachers make about classroom practices can either greatly facilitate student learning or serve as an obstacle to it.

Qualitative studies are, by their nature, in-depth portraits of the experiences of specific students and teachers. As such, they provide valuable insight into the interrelationships between various aspects of teacher practice and student learning. However, because they focus on one specific setting, it is difficult to generalize the results of these studies to broader groups of students and teachers. This suggests the need for large-scale quantitative studies that can test the

generalizability of the insights from qualitative research.

Yet there has been little quantitative research into whether classroom practices, in concert with other teacher characteristics, have an impact on student learning that is comparable in size to that from background characteristics. Two notable exceptions are a study of the classroom experiences of the nation's students using NELS:88 (National Center for Education Statistics, 1996) and a study of the professional development experiences and classroom practices of California's teachers (Cohen & Hill, 2000). The NELS:88 study related a few classroom practices to student achievement in mathematics and science and found that a focus on higher-order thinking skills had a positive effect in math but not in science. The California study related a few professional development experiences of teachers to their classroom practices, and related both of these to student scores on the state assessment. The study found positive relationships between reform-oriented classroom practices and student achievement as well as between reform-oriented professional development and reform-oriented classroom practices, although these relationships were marginal (mostly significant at the .15 level). While these two studies represent an important departure from production function studies, in their inclusion of measures of classroom practice and professional development, the usefulness of their findings is limited by their data and method. The measures of classroom practice in the NELS:88 and California databases are hardly comprehensive. Neither database has, among other things, a measure of hands-on learning activities. And the California study combines its few classroom practices into two variables, reform-minded and traditional practice, making it difficult to gauge the effectiveness of particular practices. The NELS:88 data also lack measures of most aspects of professional development, and hence professional development was not included in the NELS:88 study. The California data lack measures of social background for individual students, and hence the California study relied upon the percentage of students in the school who received a free or reduced-price lunch, a weak measure. The two studies also relied upon regression analysis, which, as shall be seen, is problematic in the study of school effects.

These two exceptions notwithstanding, quantitative research has tended to find that the effects of student background on student achievement and other outcomes far overshadows school effects. Some of the research has found no school effects at all, while other research has found effects that are, at best, modest. Specifically in terms of teaching, such research has found that most characteristics of teachers do not matter, and the few that do are not as important as student background. Yet such studies ignore qualitative work that suggests that certain classroom practices are highly conducive to student achievement. If this is the case, then classroom practices may indeed explain a substantial portion of the variance in student achievement. The current study seeks to explore this possibility, through the analysis of a national database that includes an unprecedentedly comprehensive set of classroom practices.

Hypotheses, Data and Method

The study tests two hypotheses concerning teacher quality. Teacher quality has three aspects: the teacher's classroom practices, the professional development the teacher receives in support of these practices, and characteristics of the teacher external to the classroom, such as educational attainment. The first hypothesis is that, of these aspects of teacher quality, classroom practices will have the greatest impact on student academic performance, professional development the next greatest, and teacher inputs the least. The rationale for this expectation is that the classroom is the primary venue in which students and teachers interact; hence, decisions by teachers as to what to do in this venue will most strongly affect student outcomes. Teacher inputs will be least likely to influence student academic performance because they do so less directly, through encouraging classroom practices conducive to high student performance. Professional development falls somewhere between classroom practices and teacher inputs. It does occur outside the classroom, but is more closely tied to specific classroom practices than are teacher inputs. Second, it is hypothesized that teacher quality is as strongly related to student academic performance as student background characteristics. When the effects for all three aspects of teacher quality are added together the result will be comparable in size to that of student background. The rationale behind this expectation is that, as the qualitative literature suggests, student learning is a product of the interaction between students and teachers, and both parties contribute to this interaction.

To test these hypotheses, this study makes use of NAEP, which can measure all three aspects of teacher quality as well as student performance and other potential influences on student performance. NAEP is administered every year or two in various subjects to nationally representative samples of fourth, eighth and twelfth graders. The subjects vary, but have included at one time or another mathematics, science, reading, writing, geography and history. In addition to the test itself, NAEP includes background questionnaires completed by the student, the principal, and the teacher in the relevant subject area. The results from NAEP are used to measure trends in student performance over time and to compare performance among various subgroups of students such as males and females (for an overview of NAEP, see Johnson 1994).

For this study, data on the 7,146 eighth graders who took the 1996 mathematics assessment are analyzed. Eighth graders are used for this analysis because they are exposed to a wider range of subject matter than fourth graders, and teacher questionnaires are not available for twelfth graders. Student performance is measured from test scores on the assessment. Student background is measured utilizing six questions from the student background questionnaire: the father's level of education, the mother's level of education, whether there are 25 or more books in the home, whether there is an encyclopedia in the home, whether the family subscribes to a newspaper and whether the family subscribes to a magazine. The three aspects of teacher quality are measured from a background questionnaire, completed by the mathematics teacher. Three teacher inputs are measured: the teacher's education level, whether the teacher majored or minored in the relevant subject area (mathematics or math education), and the teacher's years of experience. Ten measures of professional development are used: the amount of professional development teachers received last year

and whether teachers received any professional development in the last five years in the topics of cooperative learning, interdisciplinary instruction, higher-order thinking skills, classroom management, portfolio assessment, performance-based assessment, cultural diversity, teaching special-needs students, and teaching limited-English-proficient (LEP) students. Finally, 21 classroom practices are utilized: addressing algebra, addressing unique problems, addressing routine problems, using textbooks, using worksheets, having students talk about mathematics, having students write reports, having students solve problems that involve writing about math, having students work with objects, having students work with blocks, having students solve real-world problems, having students hold discussions in small groups, having students write a group paper, having students work with partners, assessing student progress from tests, assessing student progress from multiple-choice tests, assessing student progress from tests involving constructed responses, assessing student progress from portfolios, assessing student progress from individual projects, and the amount of homework assigned. One school characteristic not pertaining to teacher quality is also drawn from the teacher questionnaire, the number of students in the class. See Table 1 for a complete list of variables.

Table 1
Descriptive Statistics for Teacher Inputs and
Professional Development

Teacher Inputs	M	SD
Teacher's Education Level (From 1=<B.A. to 4=>M.A.)	2.38	.46
Teacher Majors in Mathematics (1=yes, 0=no)	.69	.43
Teacher's Years of Experience (From 1= 2 or less to 5=25 or more	3.53	1.17
Professional Development		
Classroom Management (1=yes, 0=no)	.44	.46
Cooperative Learning (1=yes, 0=no)	.68	.44
Cultural Diversity (1=yes, 0=no)	.32	.43
Higher-Order Thinking Skills (1=yes, 0=no)	.45	.46
Interdisciplinary Instruction (1=yes, 0=no)	.50	.47
Limited English Proficiency (1=yes, 0=no)	.12	.47
Performance-based Assessment (1=yes, 0=no)	.12	.35
Portfolio Assessment (1=yes, 0=no)	.36	.45
Special-needs Students (1=yes, 0=no)	.26	.41
Amt. Professional Development Last Year (1=none to 5=35+hours)	3.30	1.14

The method employed in this study is intended to address key methodological problems in the prior literature. Much of school effects research (including most production function studies as well as the NELS:88 and Cohen & Hill studies of classroom practice) relies upon OLS regression techniques. One problem with such techniques is that they are not sensitive to the multilevel nature of the data. School effects involve relating variables at one level of analysis, the school, to another level of analysis, the student. Studies using OLS tend either to aggregate student data to the school level or to disaggregate school data to the student level. The first approach can introduce aggregation biases into the models, the second approach can seriously underestimate standard errors, and both approaches can miss important information about the nature of the school effects (Bryk & Raudenbush, 1992; Goldstein, 1995). A second problem with regression techniques is their failure to take measurement error into account. These techniques assume that the variables in the models are perfectly measured by the observed data. Yet the operationalizations of most variables are subject to substantial error, both because the operationalization does not correspond perfectly to the model (e.g. parents' income as a proxy for socioeconomic status) and because data collection procedures are error-prone. Failing to take measurement error into account can lead to biased estimates of model coefficients. A third problem is that regression techniques are not adept at measuring interrelationships among independent variables. School effects often involve a multi-step process, in which one school characteristic influences another that may, in turn, influence the outcome of interest. While it is possible to run a series of models that regress each independent variable on the others, such models tend to be cumbersome and lack statistics measuring the overall fit of the series of models. Because of these difficulties, school effects research often neglects the indirect effects of various school characteristics.

166

One way to address these problems is through the technique of multilevel structural equation modeling (MSEM). Structural equation modeling (SEM) involves two components: factor models and path models (Hayduk, 1987; Jöreskog & Sörbom, 1993). The factor models relate a series of indicators, known as manifest variables, to a construct of those indicators, known as a latent variable. The path models then relate the latent variables to one another. The estimation procedure for both the factor and path components involves three steps. A set of hypothesized relationships is specified by the researcher. Then, through an iterative process, differences in the covariance matrix those relationships imply (Σ) and the covariance matrix of observed data (S) are minimized. The resulting estimates include coefficients for the hypothesized relationships, t-tests for their statistical significance, and statistics for the goodness of fit between Σ and S. SEM can be adapted to handle multilevel data by employing the estimation procedure separately for the two levels of analysis (Muthén, 1994; Muthén, 1991). The researcher hypothesizes a student-level factor model, a student-level path model, a school-level factor model and a school-level path model. These models can be used to generate two implied covariance matrices, Σ_B, a between-school matrix computed as the school deviations around the grand mean, and Σ_w, a within-school matrix computed as student deviations around group means. The observed data can be similarly partitioned into between- and within-school covariance matrices (S_B and S_W).

MSEMs can address the three problems in the prior literature. First, they do distinguish between schools and students; separate models are specified for each level of analysis and related to one another through a constant. Second, these models take measurement error into account in two ways. For one, the factor models explicitly measure the amount of variance in the latent variables unexplained by the manifest variables. In addition factor models can actually reduce measurement error by generating latent variables from multiple manifest variables. Third, the path models estimate interrelationships among independent variables, allowing for the estimation of indirect effects. The effect sizes and t-scores of the indirect effects are produced, as well as statistics that measure the overall goodness of fit of models that simultaneously specify these interrelationships.(Note 4)

The current study produces three MSEMs. Analyses are conducted using AMOS 3.6 (Arbuckle, 1997), a SEM software package, along with STREAMS 1.8, a pre- and post-processor that simplifies the syntax and output for multilevel models (Gustafsson & Stahl, 1997). In preparation for the preprocessor, the preexisting student-level data variable labels are reduced to six characters and missing values replaced with means for the pertinent variable. The software then aggregates the student level data to the school level, and creates both a school-level covariance matrix and a pooled matrix of residual student-level covariances.(Note 5) The first MSEM relates teacher inputs to student academic performance, taking into account student socioeconomic status (SES) and class size (see Figure 1 below). The student-level factor model generates an SES construct from the six measures of student background, and an academic performance construct from a single test score. The student-level path model simply measures the covariance between SES and student academic performance. The school-level factor model generates an SES construct from school means of the six measures of student background and an academic construct from the school mean of the single test score. In addition, class size, teachers' years of experience, educational attainment and major are constructed from individual measures that correspond to these constructs. The school-level path model treats student academic performance as a function of the other constructs.
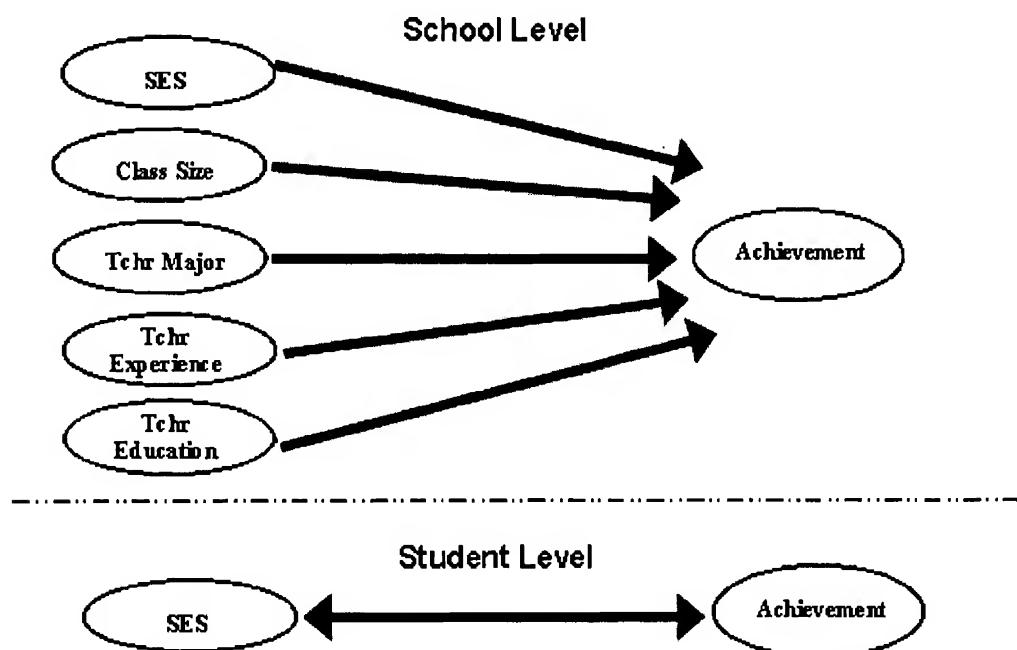

Figure 1. Teacher Inputs Path Model

167

The second MSEM relates professional development and teacher inputs to student academic performance and one another, taking into account student SES and class size (see Figure 2 below). The student-level factor and path models are the same as in the teacher inputs model. Early versions of the school-level factor and path models include SES and student academic performance, constructed as before, teacher inputs which prove significantly related to student academic performance, constructed from a single corresponding measure, the amount of time in professional development, constructed from a single corresponding measure, and all nine professional development topics. For the sake of parsimony, the final school-level factor and path models include only those professional development topics significantly related to student academic performance. These are professional development in higher-order thinking skills, constructed from a single corresponding measure, and professional development in teaching different populations of students, constructed from professional development in cultural diversity, professional development in teaching LEP students and professional development in teaching students with special needs. The parsimonious school-level path model relates each professional development construct to student achievement, and the teacher input, class size and SES both to student achievement and to each professional development construct.

School Level

Student Level
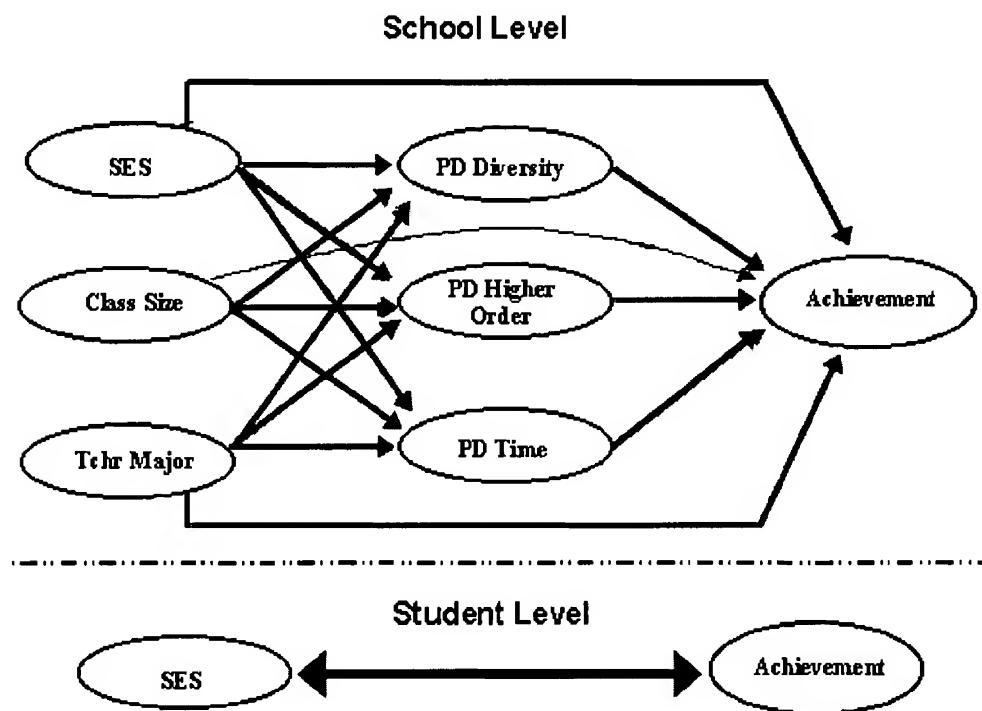
Figure 2. Professional Development Path Model

The third MSEM relates classroom practices, professional development and teacher inputs to student academic performance and one another, taking into account SES and class size (see Figure 3 below). Student-level factor and path models remain the same as in prior models. Early versions of the school-level factor and path models include SES, class size, teacher inputs that prove significant in the teacher input model, the amount of time in professional development, the topics of professional development that prove significant in the professional development model and all 21 classroom practices. For the sake of parsimony, the final school-level factor and path models include only those classroom practices that prove significantly related to student achievement. The final school-level factor model constructs the teaching of higher-order thinking skills from a single measure, solving unique problems; the teaching of lower-order thinking skills from a single measure, solving routine problems; engaging in hands-on learning from three measures, working with blocks, working with objects and solving real-world problems; assessing student progress through traditional testing from two measures, multiple choice testing and the overall frequency of testing; and assessing student progress through more authentic assessments from three measures, portfolio assessments, individual projects, and constructed response tests. The SES, class size, teacher input and professional development constructs are handled as in the professional development model. The school-level path model relates these classroom practice constructs to the student achievement constructs, relates the professional development constructs to the classroom practice constructs, and relates teacher inputs, SES and class size to the professional development classroom practice and student achievement constructs.

School Level
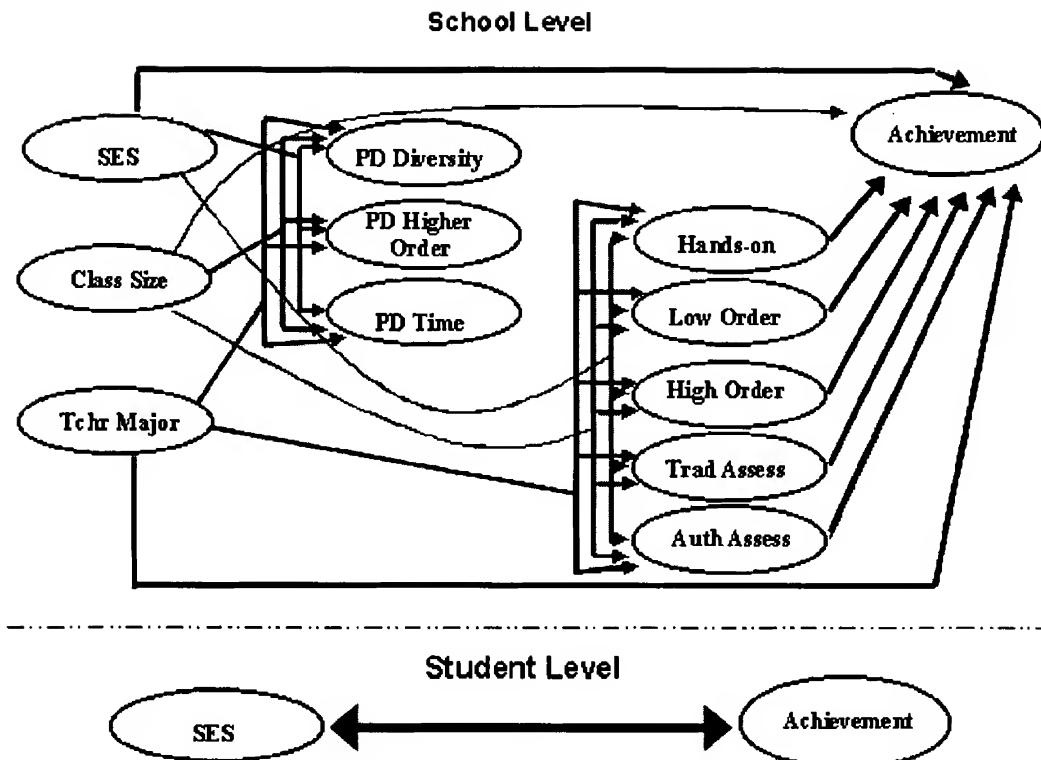
Figure 3. Classroom Practices Path Model

These procedures are modified in two ways to take the design of NAEP into account. First, design effects are employed. NAEP is a stratified, clustered sample. Secondary analyses of NAEP that treat it as a simple random sample will underestimate standard errors, making significance tests overly liberal. One procedure recommended to address this problem is to inflate standard errors estimated assuming a simple random sample by a certain factor, known as a design effect (O'Reilly, Zelenak, Rogers and Kline 1996). This study uses a design effect of 2, calculated by estimating the proper standard error for select values in the first MSEM and choosing the most conservative one. Cutoff points for all significance tests are increased by 41% (the increase in standard errors attributable to the square root of the design effect).(Note 6) Second, each MSEM is estimated multiple times, once for each "plausible value" of the student test score, and the resulting parameters and standard errors are pooled. Because each student answers only a small subset of the assessment items, it is not possible to estimate a single student score. Instead, five estimates are provided based upon the items the student did not answer and background information about the student and the school. The appropriate procedure for secondary analyses using these five estimates, which are known as plausible values, is to estimate five separate models for each of the plausible values, pool their point estimates by taking their means and pool their standard errors as the sum of the mean standard error and the variance among the five plausible values, weighted by a factor of 1.2 (Johnson, Mislevy and Thomas 1994).(Note 7) The current study employs this technique, producing a total of 15 sets of estimates, five for each of the MSEMs.(Note 8)

Results

Before discussing the results from the three MSEMs, it is worthwhile to summarize what the NAEP data reveal about the prevalence of classroom practices, professional development and teacher inputs.(Note 9) The data on teacher inputs indicates that eighth grade math teachers are most likely to possess less than a master's degree, have majored or minored in mathematics or math education, and have 10 or more years of experience teaching (Table 1). Approximately 40% of eighth graders have teachers who possess a master's degree or more, with the remainder possessing a bachelor's degree or less. Approximately 70% of eighth graders have teachers who majored or minored in mathematics or math education; the rest have teachers who are teaching off-topic. And approximately 60% of eighth graders have teachers with more than 10 years of experience.

The data on professional development indicate that while most teachers receive some professional development in some topics, that professional development tends not to be of long duration, and certain topics tend to be neglected (Table 1). Most eighth graders have teachers who received some professional development in the last five years in the most common topics, such as cooperative learning or interdisciplinary instruction. But only one-third of eighth graders have teachers who received professional development in cultural diversity, one-quarter have teachers who received professional development in teaching students with special needs, and one-tenth have teachers who received professional development in teaching LEP students. And regardless of the topic of professional development, only a

169

minority of students have teachers who received more than 15 hours of professional development last year.

The prevalence of classroom practices varies greatly (see Table 2). While much of the material covered in eighth grade involves issues of operations and measurement, teachers do cover more advanced topics. More than half of all students are exposed to algebra, and one-quarter to geometry. The kinds of problems students are taught to solve tend to involve a routine set of algorithms; four out of five students commonly work with such problems, as opposed to about half of students working with problems that involve unique situations. All students report taking a math test at least once a month. The nature of the test varies, however. Typically, students take tests that involve extended written responses (more than half do so at least once a month). About one-third of students take multiple-choice tests. Students are also assessed through individual projects and portfolios (also about one-third of students at least once a month). Hands-on learning activities appear quite infrequent. Just one-quarter of students work with objects and just one-tenth work with blocks. Problems with a concrete or practical bent, that address real-world situations, are fairly usual, however, with three-quarters of students encountering such problems as least once a week. Writing about mathematics is fairly uncommon, just one-third of students doing so at least once a week. Group activities vary in their frequency; most students discuss math in small groups, but only a minority of students solve problems in groups or work on a problem with a partner. Finally, textbooks and homework are ubiquitous in eighth-grade classrooms; nearly all students use a textbook at least once a week, and most do some homework every day.

Table 2
Descriptive Statistics for Classroom Practices

Classroom Practices	M	SD
Address Algebra (From 1=none to 4=a lot)	2.51	.59
Address Geometry (From 1=none to 4=a lot)	2.00	.61
Address Solving Routine Problems (From 1=none to 4=a lot)	2.78	.43
Address Solving Unique Problems (From 1=none to 4=a lot)	2.44	.56
Assessment Using Multiple Choice Questions (from 1=never to 4= a lot/ twice a week)	1.99	.83
Assessment Using Short/Long Answers (from 1=never to 4= a lot/twice a week)	2.49	.92
Assessment Using Portfolios (from 1=never to 4= a lot/twice a week)	1.87	.79
Assessment Using Individuals Projects (from 1=never to 4= a lot/twice a week)	2.19	.81
Work with Blocks (From 1=never to 4=almost everyday)	1.52	.58
Work with Objects (From 1=never to 4=almost everyday)	2.09	.77
Solve Real-Life Problems (From 1=never to 4=almost everyday)	2.93	.74
Write Reports (From 1=never to 4=almost everyday)	1.39	.49
Write about Math (From 1=never to 4=almost everyday)	1.97	.79
Take Math Tests (From 1=never to 4=almost everyday)	2.49	.47
Do Worksheet (From 1=never to 4=almost everyday)	2.65	.82
Talk about Math (From 1=never to 4=almost everyday)	2.70	1.02
Solve Problems with Other Students (From 1=never to 4=almost everyday)	2.84	.83
Discuss Math with Other Students (From 1=never to 4=almost everyday)	3.31	.70
Work with Partner (From 1=never to 4=almost everyday)	2.98	.82
Do Homework (From 1=never to 4=almost everyday)	2.93	.75
Use Textbooks (From 1=never to 4=almost everyday)	3.63	.65

Table 3
Descriptive Statistics for Other Characteristics of Schools and Students

Other Characteristics of Schools and Students	M	SD
Class Size (From 0=More than 36 student to 4=1 to 20 students)	2.54	.88
Student's Family Gets Newspaper (1=yes, 0=no)	.74	.43
Student's Family Has Encyclopedia (1=yes, 0=no)	.82	.38
Student's Family Gets Magazine (1=yes, 0=no)	.83	.37

BEST COPY AVAILABLE

Student's Family Has More than 25 Books (1=yes, 0=no)	.95	.21
Father's Education Level	2.91	.94
Mother's Education Level	2.85	.96
Math Score: Plausible Value #1	272.45	35.85
Math Score: Plausible Value #2	272.64	35.89
Math Score: Plausible Value #3	272.36	36.35
Math Score: Plausible Value #4	272.45	35.85
Math Score: Plausible Value #5	272.54	35.56

This description of teacher inputs, professional development and classroom practices says little about their effectiveness. The fact that certain practices are uncommon may be bad or good, depending upon their impact on student outcomes. It is the role of the series of MSEMs to gauge the effectiveness of these three aspects of teacher quality.

For all three MSEMs, the student-level factor models are similar (Table 4). The factor models show the two student-level characteristics, SES and achievement, to be well measured. All of the indicators of SES have standardized factor loadings ranging from .24 to .33, suggesting that each plays a role in constructing the variable. The construct for achievement consists of a single indicator, and hence has a loading fixed at one and an error fixed at zero. The path model consists simply of the covariance between student SES and student achievement, and this covariance proves significant, with a correlation coefficient of .35 for all models. It should be remembered that this covariance pertains only to the student-level component of the models, meaning that variations in SES among students in the same school are associated with variations in their mathematics scores within that same school. Variations in average SES and achievement between schools is the purview of the school-level models.

Table 4
Student-Level Factor and Path Models

Factor Model	Input Models			P.D. Model			Practices Model		
	SES	Ach	Err	SES	Ach	Err	SES	Ach	Err
Mother's Education Level	2.91* .31		1.00 .86	2.92* .31		1.00 .86	2.91* .31		1.00 .86
Father's Education Level	2.78* .31		1.00 .85	2.80* .31		1.00 .84	2.79* .31		1.00 .85
Family Gets Newspaper	1.00* .24		1.00 .92	1.00* .24		1.00 .92	1.00* .24		1.00 .92
Family Gets Encyclopedia	.92* .26		1.00 .94	.92* .26		1.00 .94	.92* .26		1.00 .94
Family Gets Magazine	1.16* .33		1.00 .90	1.16* .33		1.00 .90	1.16* .33		1.00 .92
Family Has More than 25 Books	.60* .30		1.00 .92	.60* .30		1.00 .92	.60* .30		1.00 .85
Plausible Value #1		1.00* .77	1.00 .28		1.00* .77	1.00 .28		1.00* .77	1.00 .28
Plausible Value #2		.99* .77	1.00 .28		.99* .77	1.00 .28		.99* .77	1.00 .28
Plausible Value #3		1.00* .77	1.00 .28		1.00* .77	1.00 .28		1.00* .77	1.00 .28
Plausible Value #4		.99* .77	1.00 .28		.99* .77	1.00 .28		.99* .77	1.00 .28
Plausible Value #5		.98* .77	1.00 .28		.98* .77	1.00 .28		.98* .77	1.00 .28
Path Model									
Covariance between SES and Achievement	1.15* .35			1.15* .35			1.15* .35		

*p<.05

The school-level factor models also have indicators that contribute substantially to their constructs (Table 5). The loadings for SES range between .17 and .25.(Note 10) Hands-on learning has loadings ranging from .46 to .79. Traditional assessment has loadings ranging from .37 to .57. And authentic assessment has loadings ranging from .41 to .73. All of the constructs generated from a single indicator have loadings fixed at 1 and errors fixed at 0 and so, by definition, their indicators contribute substantially. The one construct for which the indicators do not all contribute substantially is professional development in teaching special populations. Here two of the indicators (cultural diversity and teaching LEP students) load strongly on the construct, but the third (teaching special-needs students) does not. (A sensitivity analysis was conducted in which this indicator was excluded, without significant impact on the model.)

Table 5
School-level Factor Model: Classroom Practices

	SES	Ach	Class Size	PD Time	PD Hi Order	PD Diversity	Tchr Major	Error
Mother's Education	2.66* .23							1.00 .23
Father's Education	2.89* .25							1.00 .25
Newspaper	1.00* .19							1.00 .16
Encyclopedia	.78* .17							1.00 .07
Books	.50* .20							1.00 .10
Magazine	1.05* .23							1.00 .10
Plausible Value #1		1.00* .57						1.00 .03
Plausible Value #2		1.00* .58						1.00 .02
Plausible Value #3		1.00* .58						1.00 .03
Plausible Value #4		1.01* .58						1.00 .04
Plausible Value #5		1.00* .58						1.00 .02
Class Size			1.00* 1.00					1.00 .00
PD Time				1.00* 1.00				1.00 .00
PD Hi order					1.00* 1.00			1.00 .00
PD Cultural						1.00* .76		1.00 .46
PD LED						.65* .55		1.00 .59
PD Special						.26* .20		1.00 .69
Tchr Major							1.00* 1.00	1.00 .00

*p<.05
Cells contain unstandardized and standardized coefficients, in that order.

Table 5
School-level Factor Model: Classroom Practices (continued)

BEST COPY AVAILABLE

	Hands-On Learning	Trad Assess	Auth Assess	Lower Order	Higher Order	Error
Real-world Problems	.64* .46					1.00 .63
Work with Objects	1.00* .66					1.00 .53
Work with Blocks	.83* .79					1.00 .43
Take Tests		.35* .37				1.00 .66
Assess through Multiple-Choice Tests		1.00* .57				1.00 .58
Assess through Extended Response Tests			1.01* .65			1.00 .54
Assess through Projects			1.00* .73			1.00 .48
Assess through Portfolios			.63* .41			1.00 .65
Address Routine Problems				1.00* 1.00		1.00 .00
Address Unique Problems					1.00* 1.00	1.00 .00

*p<.05

Cells contain unstandardized and standardized coefficients, in that order.

The school-level path model for teacher inputs shows that one of the three inputs, the teacher's major, is modestly associated with academic achievement. The model consists of a single dependent variable, achievement, related to five independent variables, SES, class size and the three teacher inputs (Table 6). SES has an effect size of .76, which far overshadows those of class size and teacher's major (.10 and .09 respectively). The teacher's level of education and years of experience prove unrelated to student achievement.

Table 6
School-level Path Model: Teacher Inputs

	Ach
SES	198.41** .76
Class Size	3.04* .10
Tchr Major	4.82** .09
Tchr Ed	1.20 .02
Tchr Exp	1.03 .05
Error	1.00 .44
*p<.10 **p<.05	

The school-level path model for professional development finds that two topics, addressing special populations of students and higher-order thinking skills, are substantially related to student achievement. The model indicates that schools with high percentages of affluent students tend to have less time spent on professional development generally, and are less likely to expose their teachers to professional development on working with different student populations (Table 7). Schools with smaller average class sizes are also less likely to do these things. But, schools with more

teachers teaching on topic also devote more time to professional development. Of the three aspects of professional development, the amount of time is not significantly related to achievement. Professional development in higher-order thinking skills, and dealing with special populations, however, do have significant effects, with standardized coefficients of .12 and .21 respectively.

Table 7
School-level Path Model: Professional Development

	PD Diversity	PD Hi Order	PD Time	Ach
SES	-1.29** -.32	-.58 -.09	-1.87* -.12	213.18** .83
Class Size	-.08** -.17	-.04 -.06	-.20* -.11	4.23** .14
Tchr Major	-.01 -.01	.11 .08	.94** .27	5.05** .09
PD Diversity				13.24** .21
PD Hi order				4.88** .12
PD Time				-.23 -.01
Error	1.00 .66	1.00 .70	1.00 .67	1.00 .41
*p<.10 **p<.05 Cells contain unstandardized and standardized coefficients, in that order.				

The school-level path model for classroom practices finds three constructs, hands-on learning, solving unique problems and avoiding reliance on authentic assessments, to be positively related to student achievement (Table 8). All five of the classroom practice constructs are related to some of the earlier variables, SES, class size, teacher major or the three aspects of professional development. Schools with more affluent students are more likely to solve unique problems and less likely to engage in inauthentic forms of assessment. Schools where teachers received professional development in dealing with different student populations are less likely to have students engage in routine problem solving. And schools where teachers received professional development in higher-order thinking skills are more likely to have students engage in hands-on learning. Also, the more time teachers engage in professional development, the more their students engage in hands-on learning and authentic assessment. These practices are associated with student achievement. Schools where students engage in hands-on learning score higher on the mathematics assessment. Schools where students solve unique problems also score higher, as do those schools that do not rely primarily on authentic forms of assessment.

Table 8
School-level Path Model: Classroom Practices

	PD Diversity	PD Hi Order	PD Time	Hands-On Learning	Lower Order	Higher Order	Trad Assess	Auth Assess	Ach
SES	-1.11** -.05	-.43 -.07	-1.83* -.12	-1.02 .14	-.32 -.06	1.15** .17	-2.35** -.34	-.79 -.09	192.26** .74
Class Size	-.09** -.17	-.04 -.06	-.20* -.11	.03 .03	.03 .05	.03 .03	.01 .01	-.09 -.10	2.33 .08
Tchr Major	-.01 -.01	.11 .08	.94** .27	-.04 -.02	-.06 -.05	.02 .01	.19 .13	-.05 -.03	4.19* .07
PD Diversity				-.23 -.14	-.30* -.24	-.24 -.16	-.13 -.09	-.14 -.07	
PD Hi order				.34** .30	.01 .01	.12 .11	.21 .19	.23* .18	

PD Time				.13** .27	.03 .08	.05 .12	-.14** -.32	.14** .26	
Hands-On Learning									8.88** .25
Lower Order									-3.85 -.08
Higher Order									4.82** .13
Trad Assess									1.23 .03
Auth Assess									-5.73** -.18
Error	1.00 .67	1.00 .70	1.00 .67	1.00 .63	1.00 .68	1.00 .68	1.00 .62	1.00 .66	1.00 .40

*p<.10
**p<.05
Cells contain unstandardized and standardized coefficients, in that order.

Comparisons among the three school-level path models help to gauge the impact of teaching on student achievement. First, all of the models explain a similar amount of variance. While the residual variance goes from .44 in the teaching model to .41 in the professional development model and .40 in the classroom practices model, these differences are slight. Thus, rather than explain more variance, the more complex models simply reallocate variance among explanatory variables. Second, the three models show the total effect of each teacher quality variable. The total effect is the sum of all direct and indirect effects, and is measured for each aspect from the sum of the effect sizes of the variables in directing that aspect in the model in which that aspect is related to achievement without mediating variables.(Note 11) Thus, the effect size of the one significant teacher input is .09, taken from the teacher inputs model; the effect sizes for the statistically significant aspects of professional development total .33, taken from the professional development model; and the effect sizes for the classroom practices total .56, taken from the classroom practices model. Third, all of the models fit the data well, with goodness of fit indices at the .99 or 1.00 level and root mean squared errors of approximation at the .014 level or better.

In sum, it appears that the various aspects of teacher quality are related to student achievement when class size and SES are taken into account. In particular, the following 5 variables are positively associated with achievement:

- Teacher major
- Professional development in higher-order thinking skills
- Professional development in diversity
- Hand-on learning
- Higher-order thinking skills

Before discussing further the implications of these results, however, it is necessary to note some shortcomings of the study.

Methodological Caveats

The study suffers from four basic shortcomings. First, the data are cross-sectional. The information about aspects of teacher quality is collected at the same time as student test scores. Consequently, it is not possible to draw inferences about the direction of causation for the relationships that were discovered. It may be that a focus on higher-order thinking skills causes increased student performance, or it may be that having high-performing students drives teachers to focus on higher-order thinking skills. The likelihood of the latter scenario is somewhat reduced in that the models take SES and class size, both proxies of prior academic performance of the student and school, into account. Nonetheless, to confirm the causal direction hypothesized in this study, subsequent research should replicate the results using longitudinal data.

Second, the study covers only one grade level in one subject. It is possible that different sets of classroom practices will prove effective for other subjects and at other grade levels. Third, this study does not measure the link between aspects of teacher quality and the relationship between student test scores and student SES. MSEM measures student-level covariances by pooling each school's within-school covariance matrix. Consequently, while it is possible to measure the relationship between a school variable and a student outcome, it is not possible to measure the relationship between a school variable and the relationship between two student characteristics. Other multilevel techniques, such as Hierarchical Linear Modeling, while unable to perform certain analyses that MSEM can perform (e.g. confirmatory factor analysis), are able to accomplish this. Subsequent research should supplement the findings of

this study by measuring the impact of classroom practices and other aspects of teacher quality on the relationship between student test scores and student background characteristics. Such analyses will make it possible to know not only how teachers can affect the average performance of their class, but how they can affect the distribution of performance within the class.

Finally, better indicators of the constructs used in this study are needed. The SES construct lacks indicators of parents' income or occupation, as well as non-educational materials in the home such as a microwave or washer and dryer, indicators which prior research has found to be an important component of SES. Exposure to each topic of professional development is measured as whether the teacher had received any exposure in the last five years, making it impossible to distinguish between professional development that is rich and sustained and a lone weekend seminar. It is also not possible to measure how receptive teachers are to the professional development they receive. Presumably, a more attentive teacher would benefit more from professional development than a less attentive one. Given that professional development in working with different student populations is so important, it would be useful to include a measure of classroom practices that involves this activity. And while many of the classroom practices are measured through multiple indicators, some, such as higher-order thinking skills, are not. Additional indicators for single-indicator constructs should be introduced to increase the reliability of the constructs.(Note 12)

Conclusion

Despite these methodological shortcomings, the current study represents an advance over previous work. The first model to some extent exemplifies the traditional approach to gauging the impact of teaching and other school characteristics on student achievement. Although the model differs from most production function studies in including a measurement component and being multilevel, it is otherwise similar. Like OLS, the model relates a single dependent variable to a series of independent variables. The independent variables consist of teacher inputs and a class size measure, controlling for student background. Like most of the prior research, this model finds no significant relationship to test scores for most of the characteristics, with the exception of the teacher's college-level coursework as measured by major or minor in the relevant field. And like all of the prior research, all school effects are overshadowed by the effect of student SES.

The subsequent models move beyond the first by introducing measures of what teachers actually do in the classroom, the training they receive to support these practices directly, and by modeling interrelationships among the independent variables. They are able to do so because the NAEP database includes a comprehensive set of classroom practices, and because MSEM can model all of the relevant interrelationships. And all of the models, including the teacher inputs model, move beyond most prior research in their ability to take into account measurement error and the multilevel nature of the data. Through these innovations it was possible to confirm the two hypotheses regarding the role that teaching plays in student learning.

The first hypothesis, that, of the aspects of teacher quality, classroom practices will have the greatest effect, is confirmed by the models. The effect sizes for the various classroom practices total .56; those for the professional development topics total .33; and the effect size for the one teacher input found to have a statistically significant impact is .09. As the qualitative literature leads one to expect, a focus on higher-order thinking skills is associated with improved student performance. Applying problem-solving techniques to unique problems is a key component of such skills. Hands-on learning can be understood in this way as well, in that it involves the simulation of concepts, moving the student from the abstract to the concrete.(Note 13) Also suggested by the qualitative literature, individualizing instruction seems to be effective. Students whose teachers received professional development in learning how to teach different groups of students substantially outperformed other students. One apparent inconsistency between the findings of this study and the qualitative literature is in the area of authentic assessment in that the study documents the importance of using some form of traditional testing in assessing student progress. This finding, however, merely suggests that on-going assessments such as portfolios and projects are not sufficient; they need to be supplemented with tests that occur at a distinct point in time.

The second hypothesis, that the total impact of the teaching variables will be comparable to that of student SES, is also confirmed. The sum of the effects from the three aspects of teacher quality is .98. The effect sizes for SES range from .74 to .83, with a value of .76 in the model where all three aspects of teacher quality are included (the classroom practices model). Thus, the impact of teaching can be said not only to be comparable to that of SES, but even to be somewhat greater.

In addition to confirming the hypotheses regarding the impact of teaching on student learning, the study uncovers important interrelationships among the aspects of teaching. For one, professional development seems to influence teachers' classroom practices strongly. The more professional development teachers received in hands-on learning, and indeed the more professional development they received regardless of topic, the more likely they are to engage in hands-on learning activities. And the more professional development teachers received in working with special student populations, the less likely they are to engage in lower-order activities. Another important interrelationship involves the trade-off between teacher quality and teacher quantity. Smaller class sizes are negatively associated with teachers majoring in their relevant subject and in receiving substantial amounts of professional development, whereas teacher major and time in professional development are positively associated with one another. These relationships suggest that schools tend to choose between hiring more teachers or investing in improved teacher quality through recruiting teachers with better preservice training and providing teachers with more and better in-service training.

176

In sum, this study finds that schools matter because they provide a platform for active, as opposed to passive, teachers. Passive teachers are those who leave students to perform as well as their own resources will allow; active teachers press all students to grow regardless of their backgrounds. Passive teaching involves reducing eighth-grade mathematics to its simplest components. All lessons are at a similar level of abstraction; problems are solved in a single step and admit of a single solution; and all students are treated as if they had entered the class with the same level of preparation and the same learning styles. In contrast, active teaching does justice to the complexities of eighth-grade mathematics. Lessons work at multiple levels of abstraction, from the most mundane problem to the most general theorem; problems involve multiple steps and allow multiple paths to their solution; and teachers tailor their methods to the knowledge and experience of each individual student. Schools that lack a critical mass of active teachers may indeed not matter much; their students will be no less or more able to meet high academic standards than their talents and home resources will allow. But schools that do have a critical mass of active teachers can actually provide a value-added; they can help their students reach higher levels of academic performance than those students otherwise would reach. Through their teachers, then, schools can be the key mechanism for helping students meet high standards.

Acknowledgements

Notes

[1] As is common in the literature, this articler uses the terms "effect" and "school effect" to connote statistically significant associations between variables. These associations need not be causal in nature.

[2] For a discussion of the methodological issues associated with production function research, see Wenglinsky (1997), Forture and O'Neil (1994) and Monk (1992).

[3] For mathematics, the classroom practices are similar to those endorsed by the National Council on Teaching Mathematics (1989).

[4] It should be noted that some school effects research addresses the problem of the insensitivity of regression analysis to multilevel data through the use of hierarchical linear modeling (HLM). There are trade-offs to using HLM as opposed to MSEM. HLM has the advantage of being able to treat as a dependent variable not only a student outcome, but the relationship between that outcome and a student background characteristics; for its part, MSEM makes it possible to explicitly model measurement error and more fully test relationships among independent variables. While this study uses MSEM, it should be supplemented with an HLM.

[5] In aggregating teacher characteristics to the school level, the values of all teachers in that school for whom there were data were averaged. It was not possible to create a separate teacher level of analysis because there were generally only one or two teachers surveyed from each school, and thus not a sufficient number of degrees of freedom for a third level.

[6] For a fuller discussion of this approach as applied to the 1992 mathematics assessment for eighth graders, see Wenglinsky (1996).

[7] More generally, the pooled variance can be expressed as:
$$V = U^* + (1 + M^{-1})B,$$
Where V is the pooled variance,
U^* is the average sampling variance,
M is the number of plausible values,
and B is the variance among the M plausible values.

[8] One misleadingly compelling alternative to this approach is to treat the five plausible values as multiple indicators of a test score construct. However, this approach violates the assumption in structural equation models of independence of errors, and has been shown to distort estimates of residual variances and certain statistics, such as the R-squared (Mislevy 1993).

[9] Because NAEP is a sample of students and schools, but not of teachers, descriptive statistics apply to the students rather than the teachers (e.g. 45% of students have teachers who received professional development in higher-order thinking skills, not 45% of teachers received professional development in higher-order thinking skills).

[10] Loadings used here are taken from the classroom practices model (Table 5). For constructs that were also included in other models, the loadings proved nearly identical across models. The output for the two other school-level factor models is not presented have but is available upon request.

[11] Total effects can be calculated in one of two ways. The first is to estimate a single model that includes all relevant variables, both exogenous and endogenous, and to sum each of the direct and indirect effects for each variable. This option can be problematic, however, in that the size of the total effect may be an artifact of the number of paths the model permits. The more paths that are fixed at zero, for a given variable, the lower the total effect. The second option is to estimate successive models, in which only the direct effects of the variables are used. Thus, in the current case, the first model is made entirely of exogenous variables. Their direct effects on achievement are equal to their total effects. The second model adds a set of endogenous variables. They are related to achievement only in a direct manner, however, and hence can be treated as total effects. A final set of endogenous variables is added in the third model. These, too, are only directly related to achievement and hence can be treated as estimates of total effects. The presentation of total effects in this study is thus based upon the direct effects of teacher inputs in the first model, of professional development in the second model, and of classroom practices in the third model.

[12] Mayer (1999) finds that while composite measures of classroom practices drawn from teacher questionnaires are highly reliable and valid, individual measures are problematic.

[13] That said, hands-on learning may not always tap higher-order thinking skills. If a teacher does not make use of hands-on activities in a manner that connects them to underlying concepts, these activities may degenerate into a set of cookbook procedures. The fact that, as this study suggests, it is the better-trained teachers who utilize hands-on techniques suggests, however, that such connections do tend to be made.

References

Arbuckle, J. L. (1997). *Amos users' guide. Version 3.6*. Chicago: Small Waters Corporation.

Austin, G. R. & Garber, H. (Eds.). 1985. *Research on exemplary schools*. New York: Academic Press.

Brookover, W., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *School social systems and student achievement: Schools can make a difference*. Brooklyn, NY: J. F. Bergin.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Chubb, J. & Moe, T. (1990). *Politics, markets and America's schools*. Washington, D.C.: The Brookings Institution.

Cohen, D. K. & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294-343.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York. R. L. (1966). *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-24.

Ehrenberg, R. G. & Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. *Economics of Education Review, 14*(1), 1-21.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation, 28*(2), 465-498.

Ferguson, R. F. & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding school accountable: Performance-based reform in education* (pp. 265-298). Washington, D.C.: The Brookings Institution.

Goldhaber, D. D. & Brewer, D. J. (1996) Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32*(3), 505-520.

Goldstein, H., (1995). *Multilevel statistical models (2nd ed.)* New York: Halsted Press.

Golub, J. (Ed.) (1988). *Focus on collaborative learning*. Urbana, IL: National Council of Teachers of English.

Graves, D. H. & Sunstein, B. S. (Eds.) (1992). *Portfolio portraits.* Portsmouth, NH: Heinemann.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Gustafsson, J. E. & Stahl, P. A. (1997). *STREAMS user's guide (Version 1.7).* Mölndal, Sweden: Multivariate Ware.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Research, 18*(4), 45-51.

Hanushek, E. A. (1996a). A more complete picture of school resource policies. *Review of Educational Research, 66* (3), 397-409.

Hanushek, E. A. (1996b). School resources and student performance. In G. T. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43-73). Washington, D.C.: The Brookings Institution.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins University Press.

Hedges, L. W. & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G. T. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 74-92). Washington, D.C.: The Brookings Institution.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Research, 23*(3), 5-14.

Jencks, C., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.

Johnson, E. (1994). Overview of part I: The design and implementation of the 1992 NAEP. In E. Johnson & J. Carlson, *The NAEP 1992 Technical Report* (pp. 9-32). Princeton, NJ: Educational Testing Service.

Johnson, E., Mislevy, R. J., & Thomas, N. (1994). Scaling procedures. In E. Johnson & J. Carlson, *The NAEP 1992 Technical Report* (pp. 241-256). Princeton, NJ: Educational Testing Service.

Jöreskog, K. G. & Sörbom, D. (1993). *Structural equation modeling and the SIMPLIS command language.* Chicago: Scientific Software International.

Langer, J. A. & Applebee, A. N. (1987). *How writing shapes thinking.* Urbana, IL: National Council of Teachers of English.

Lee, V. E., Bryk, A. S., & Smith, J. B. (1993). The organization of effective secondary schools. *Review of Research in Education, 19*, 171-267.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29-45.

McLaughlin, M. E. & Talbert, J. E. (1993). Introduction: New visions of teaching. In M. W. McLaughlin & J. E. Talbert (Eds.), *Teaching for understanding* (pp. 1-10). San Francisco, CA: Jossey-Bass.

Mehan, H. (1992). Understanding inequality in schools: The contribution of interpretive studies. *Sociology of Education, 65*(1), 1-20.

Monk, D. H. (1992). Educational productivity research: An update and assessment of its role in education finance reform. *Educational Evaluation and Policy Analysis, 14*, 307-332.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125-145.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 399-420.

National Center for Education Statistics. (1996). *High school seniors' instructional experiences in science and mathematics.* Washington, D.C.: U.S Government Printing Office.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.*Reston, VA: Author.

O'Reilly, P. E., Zelenak, C. A., Rogers, A. M., & Kline, D. L. (1996). *1994 trial state assessment program in reading secondary-use data files user guide.*Washington, D.C.: U.S. Department of Education.

Steinberg, L. D. (1996). *Beyond the classroom: Why school reform has failed and what parents need to do.* New York: Simon and Schuster.

Strauss, R. P. & Sawyer, E. A. (1986). Some new evidence on student and teacher competencies. *Economics of Education Review,*5(1), 41-48.

Traub, J. (2000, January 16). What no school can do. *New York Times Magazine,*p. 52.

Wenglinsky, H. (1997). How money matters: The effect of school district spending on academic achievement. *Sociology of Education, 70*(3), 241-256 221-237.

About the Author

Harold Wenglinsky
Email: hwenglinsky@msn.com

Dr. Harold Wenglinsky worked for six years with the Policy Information Center at Educational Testing Service as a National Assessment of Educational Progress Visiting Scholar, Research Scientist, and Center Director. His research has tackled a wide range of issues of educational policy, from the impact of educational expenditures on student performance to the educational role of Historically Black Colleges and Universities. His most recent publications have focused on the issue of teacher quality and, more specifically, the link between various teaching practices and student outcomes, as well as identifying methodologies appropriate for using large-scale data to study such effects. He is currently writing a book on the impact of educational technology on teaching and learning.

Los Rituales Escolares y las Prácticas Educativas

Magister Pablo Daniel Vain
Universidad Nacional de Misiones
Argentina

Abstract
This study on school rituals, based on an socio-anthropology view, has arisen from the hypothesis of the anthropologist Roberto Da Matta. This hypothesis supports the theory that rituals are useful, particularly in a complex society, to promote its social identity and develop its character. Da Matta observes that it is as if the ritual domain were a privileged area from whence to enter the cultural kernel of a society, its main ideology, its system of values. This is the reason why we have put forward the proposal that to enquire about rituals at schools can result in a useful contribution to the analysis of this institution in its reproductive dimension or in the construction of a determined social structure. This research was carried out in three schools in the city of Posadas, Misiones, Argentina. In two of them, the research was pursued as a sustained, long-term and ethnographic observation: students, parents, teachers and the managing staff were interviewed. In the third school, just the teachers and the managing staff were interviewed by means of a probing survey; in both cases, strategies, sources and techniques were combined.

Resumen
En este estudio, (Nota 1) de carácter socioantropológico, sobre los rituales escolares se ha partido de la hipótesis del antropólogo Roberto Da Matta quién sostiene que "...los rituales sirven, sobre todo en la sociedad compleja, para promover la identidad social y construir su carácter". (...) " Es como si el dominio del ritual fuese una región privilegiada para penetrar en el corazón cultural de una sociedad, en su ideología dominante, en su sistema de valores.." (Nota 2) Por ello nos hemos planteado que indagar acerca de los rituales en la escuela, puede resultar un aporte interesante para analizar esta institución en su dimensión reproductora o de construcción de una determinada estructura social. La investigación se realizó en tres escuelas de la Ciudad de Posadas (Misiones; Argentina). En dos de estos establecimientos se realizó un trabajo de observación sostenida y prolongada de carácter etnográfico, se realizaron entrevistas a alumnos, docentes, directivos y padres; mientras que en la tercera de ellas solo se efectuaron entrevistas y se aplicó una encuesta-sondeo al personal docente y directivo, a modo de triangulación de estrategias, fuentes y técnicas.

1. Introducción: Rituales, prácticas y la escuela invisible

En este trabajo intentamos aproximarnos a las tramas ocultas que se generan en las instituciones educativas. La intención de desnudar mecanismos ocultos en la escuela apunta a ofrecer a los educadores un material de reflexión sobre su práctica profesional. En la Argentina, es relativamente sencillo observar que la institución escolar está

abarrotada de rituales. Desde aquellos muy estructurados, como los "actos escolares", hasta formas ritualizadas que atraviesan lo cotidiano, como las formaciones, el saludo a las autoridades, los premios y castigos, etc.. Niños y jóvenes son expuestos a un conjunto de conductas estereotipadas y generalmente transmitidas de un modo repetitivo, y en apariencia carente de significación para ellos. Los docentes son los encargados de esa transmisión, que entendemos se produce, en la mayoría de los casos, de un modo rutinario, tradicional e inconsciente.

Por otra parte, se percibe un discurso contradictorio respecto a la escuela como institución cargada de ideología. Desde el retorno de la Argentina al sistema democrático, hace más de quince años, se escucha permanentemente hablar de una educación para la democracia, respetuosa de los derechos humanos; sin embargo, las prácticas educativas aparecen profundamente sesgadas por formas autoritarias.

Existen interesantes trabajos que han abordado el problema del autoritarismo en la educación en nuestro país (Filmus, 1988; Tedesco, Braslavsky y Carciofi, 1987), los cuales se refieren al pasado gobierno militar. Pero a nosotros nos interesa sostener que la estructura autoritaria persiste, y que no sólo proviene de la educación del período 1976-1983, sino que es el reflejo de una sociedad con una importante carga de autoritarismo, relativamente independiente de la alternancia entre gobiernos militares y civiles que ha caracterizado a los últimos cien años de nuestra historia.

Entendemos que los rituales son prácticas que intentan reproducir la estructura social a través de la reproducción de la ideología dominante; y su estudio nos proporcionó precisiones respecto a cuál es la ideología que transmite la escuela; más allá de sus discursos acerca de los valores democráticos. Las hipótesis principales de este trabajo sostienen que: por una parte, más allá de las formulaciones discursivas acerca de la democratización de la educación, la escuela transmite una ideología autoritaria; y que esto se expresa a través de los rituales escolares. Por otra parte, que los educadores transmiten esa ideología de un modo, por lo general, inconsciente.

2. Marco teórico-contextual

En este trabajo hemos abordado la relación educación y sociedad desde un enfoque multirreferenciado (Frigerio; 1995). Esta perspectiva epistemológica requiere tomar la precaución de precisar claramente desde dónde hablamos; qué consideramos es lo compatible; y qué incorporamos a nuestro análisis de cada una de las líneas teóricas que nos nutren.

La educación como práctica social

La educación es una práctica social y, en ese sentido, construida desde lo singular, lo social, lo histórico y lo político (Kemmis; 1990). En tanto práctica social está inscrita en una totalidad histórico-social (Zemelman; 1987) que le da sentido, pero es preciso realizar la siguiente advertencia: el concepto de totalidad no es tomado aquí como la idea de una estructura que se impone a los sujetos como una cosa, como un aparato. Ni en el sentido de totalidad homogénea, sino como una totalidad construida en múltiples articulaciones. Lo social se materializa en las prácticas, pero dichas prácticas ocultan relaciones sociales que deben ser puestas al descubierto. En ese sentido, la idea clave es la de "desfetichización". Desfetichizar lo social y lo educativo supone desmontar la idea que presenta a las prácticas como simples actividades; ya que las prácticas sociales y educativas esconden las relaciones sociales que las producen. En este contexto es posible afirmar que la ideología se materializa en las prácticas y no -como proponen algunos autores- que la misma tiene existencia material.

En el concepto de "habitus" propuesto por Pierre Bourdieu encontramos una noción mediadora entre estructura y acción (Bourdieu; 1991), así resulta posible pensar el doble carácter de lo social: lo social hecho cosas (o estructuras sociales externas) y lo social hecho cuerpo (o estructuras sociales internalizadas), y las articulaciones entre estas dos condiciones. Al analizar las estructuras sociales externas rescatamos, de un modo crítico, el concepto de aparato ideológico del Estado propuesto por Althusser (Althusser; 1988), señalando cómo la escuela, convertida en aparato ideológico escolar, interviene como productora de sujetos sociales, es decir como generadora de determinados habitus inculcados a partir de lo social hecho cosas. Los aparatos son entonces productores de prácticas y sujetos.

Al percibir a la educación como aparato ideológico del Estado, pensamos que estos aparatos - cuya función es originariamente reproductiva - quedan atrapados en la lucha por la hegemonía que entablan los diferentes grupos y sectores de una sociedad. Es allí donde la reproducción mecánica se hace imposible, en tanto esa lucha es productora de contradicciones y fisuras que crean intersticios en los cuales se hace posible construir alternativas al pensamiento dominante. Desde los aparatos ideológicos, la ideología dominante constituye un arbitrario cultural (Bourdieu y Passerón; 1981) que logra imponerse con especial eficacia, en tanto esconde su carácter arbitrario y se presenta como una expresión natural. Este mecanismo implica el ejercicio de una forma particular de violencia, una violencia escondida, solapada y disimulada: violencia simbólica (Bourdieu y Passerón; 1981). Los actores sociales de la práctica educativa no son conscientes de su carácter de portadores de una ideología, entendida como arbitrariedad que esconde determinado tipo de relaciones sociales. Centrándonos particularmente en los docentes y directivos de las instituciones escolares, proponemos que la eficacia de la violencia simbólica -al presentar lo arbitrario como natural, escondiendo las relaciones sociales que producen las prácticas y las representaciones- logra operar en los actores y a través de ellos.

Consideramos que la ideología se internaliza en un proceso dinámico, no se trata simplemente de una acción

mecánica de inculcación que es incorporada pasivamente por el sujeto, porque el mecanismo de imposición no deviene de un todo consolidado y homogéneo, sino que es producto de las luchas por la dominación. Tampoco el arbitrario es asimilado pasivamente por los sujetos; éstos reelaboran el arbitrario conforme a sus variables individuales, sociales, históricas y políticas; y en muchos casos tal reelaboración puede resultar contradictoria.

Al abordar el problema del poder indicamos que este no solo acciona desde la estructura social externa, sino que lo hace, también, a partir de mecanismos relacionales y cotidianos en los que el poder se construye, se desarrolla y se ejerce (Foucault; 1992). Así, el poder está en lo social hecho cuerpo y lo que define una relación de poder es que supone un modo de acción, que no actúa de manera directa e inmediata sobre los otros, sino que actúa sobre sus acciones: una acción sobre la acción, sobre acciones actuales o eventuales, presentes o futuras. (Foucault; 1990). Nuestra preocupación por el poder es, entonces, respecto a sus efectos, en tanto circula a través de prácticas institucionales y discursos de la vida cotidiana.

El dominio del ritual

A partir de esta doble manera de acción del poder (desde lo estructural y desde lo intersubjetivo) establecemos un puente para interpretar el lugar de los rituales en el contexto de las prácticas educativas. Para ello señalamos que el ritual tiene un tilde distintivo que es la dramatización, entendida como condensación de algún aspecto, elemento o relación que es focalizado o destacado (Da Matta; 1983). Describimos entonces a los rituales como asociaciones de símbolos que poseen un carácter inherentemente dramático y que comunican clasificando la información en diferentes contextos. Aquello que el ritual remarca es un elemento significativo de una cultura determinada. Como acción simbólica, el ritual subraya, destaca, resalta y torna especial cualquier acción cotidiana. Esto último implica que no hay acciones esencialmente rituales, todo puede ser ritualizado si se lo convierte en condensación de algún rasgo significativo de una cultura determinada. En consecuencia, aunque los rituales poseen una base material (suponen un espacio, tiempo, objetos y acciones determinadas, tal como el caso de los gestos) su esencia es predominantemente simbólica; en este sentido su función es poner en acto un significado. Por ello, el ritual opera en el campo de las representaciones sociales y resulta un mecanismo transmisor de ideología.

Coincidimos con Margulis (Margulis; 1994) en que somos poseedores de signos, los que, elaborados a lo largo del tiempo y en el interior de una cultura, orientan nuestra actuación. Los signos implican una construcción del mundo, una clasificación; agrupan y catalogan la inmensa diversidad que nos presenta el mundo. En este marco, los rituales hacen posible la generación de sentidos, son productores de representaciones y las representaciones orientan la formación de los habitus. Dicho de otro modo, las representaciones son mediaciones entre los contenidos del ritual y la formación de los habitus.

En función de este concepto, resulta posible pensar que los rituales son mecanismos generadores de habitus, y que los primeros operan sobre las representaciones de los sujetos, los que adquieren disposiciones duraderas para la acción. También resulta posible afirmar que la internalización de los contenidos de los rituales, en forma de representaciones sociales, producirá los habitus que un sujeto dado pondrá en juego durante la vida social. Es decir que un determinado actor social tendrá mayor o menor disposición a actuar de determinadas maneras, en función de los habitus que ha conformado a partir de las representaciones sociales que ha adquirido mediante diferentes mecanismos, entre ellos el ritual.

Los rituales tienen un aspecto político ya que pueden incorporar y transmitir ciertas ideologías o visiones del mundo o, en su carácter impugnador, puede invertir las normas y valores del orden social dominante. Tomando este lado político, y partiendo del modo en que hemos conceptualizado a la totalidad social y a la ideología, es factible proponer que -tal como afirma McLaren- que los rituales, en tanto formas actuadas de significado, posibilitan que los actores sociales enmarquen, negocien y articulen su existencia como seres sociales y culturales (McLaren; 1995). En estos procesos de negociación, no encontramos acciones rituales esencialmente narcotizantes o impugnadoras; debemos analizar en cada caso su sentido para los protagonistas y los espectadores, según los contextos o coyunturas (García Canclini; 1982).

El contexto socio-educativo

El contexto en el cual se ha desarrollado esta investigación está signado por una contradicción. Por una parte, aparece la llamada desde el gobierno "Transformación Educativa"; mientras que, por el otro, esta reforma se sitúa en el marco de un Estado que tiende a retirarse de las políticas sociales, a desentenderse de lo que se consideran gastos innecesarios para el aparato estatal, y que colocan a la educación en un proceso global de achicamiento de ese mismo Estado. Algunas políticas compensatorias (por ejemplo el Plan Social Educativo) fueron impulsadas para paliar el déficit educativo. Sin embargo, las propias autoridades nacionales las han reconocido como poco eficaces, al analizar los resultados de los Operativos Nacionales de Evaluación de la Calidad Educativa. En su interesante trabajo sobre la relación entre educación y pobreza Tenti Fanfani afirma como el empobrecimiento creciente de la población se da cita con el empobrecimiento de la educación pública nacional, en términos del deterioro de la cantidad y calidad de los recursos de la oferta educativa (Tenti Fanfani; 1992).

3. Metodología.

184

El diseño metodológico que utilizamos en esta investigación implica:

a. Sistema de hipótesis

En esta investigación, y siguiendo los supuestos epistemológicos en los cuales se fundamenta, designamos con el nombre de hipótesis las suposiciones referidas a los aspectos mencionados (Briones; 1982) Pero no utilizamos, en este caso, un sistema de hipótesis causa-efecto, sino que dadas las características de las proposiciones realizadas estas son hipótesis descriptivas en tanto se refieren a: "...la existencia, la estructura, el funcionamiento, las relaciones y los cambios de cierto fenómeno. (Briones; 1982).

Hemos formulado un sistema tentativo de hipótesis, asumiendo que el mismo sería muy probablemente incompleto; que las hipótesis serían absolutamente provisionales y que iríamos a reformulándolas en el propio desarrollo de la investigación. En esa dirección, este trabajo parte de una hipótesis principal que entiende a los rituales escolares como modos de transmisión de la ideología dominante de una sociedad.

De esta hipótesis general derivamos las siguientes:

- Los rituales escolares son transmitidos en la institución escolar por los docentes, quienes no son conscientes de que ideología transmiten.

- Los rituales escolares, en nuestras escuelas, transmiten fundamentalmente principios de un orden autoritario.

- Los rituales escolares no logran transmitir la totalidad de la ideología dominante porque la lucha por la hegemonía produce fisuras en el discurso dominante.

- Los rituales escolares no logran imponer en su totalidad la ideología dominante porque, a pesar de la efectividad de la violencia simbólica que esta utiliza para imponerse, el papel activo de los actores principales produce rupturas en la homogeneidad de su discurso.

b. Estudios de caso

La investigación se realizó en tres escuelas de la Ciudad de Posadas (capital provincial; 254.951 habitantes), que en virtud de los acuerdos efectuados con los Directivos y Docentes de las mismas serán tenidas en el anonimato. En dos de estos establecimientos se realizó el trabajo de observación sostenida y prolongada, se realizaron entrevistas a alumnos, docentes, directivos y padres; mientras que en la tercera de ellas sólo se efectuaron entrevistas y se aplicó una encuesta-sondeo al personal docente y directivo.

Las principales características de estas instituciones son:

Cuadro I
Instituciones en las que se realizará
la investigación y sus principales características.

	ESCUELAS		
CARACTERÍSTICAS	**Provincial N° 377**	**Santa Rosa de Lima**	**Provincial N° 202**
Localización	Urbana, periférica.	Urbana, centro.	Urbana, periférica
Tipo de gestión	Estatal, provincial.	Privada, religiosa.	Estatal, provincial
Matrícula	530 alumnos.	980 alumnos.	620 alumnos.
Cantidad de grados	19 secciones.	21 secciones	25 secciones
Personal docente	30 docentes.	53 docentes.	32 docentes
Nivel Económico-social predominante en el alunado.	Medio-bajo y bajo	Medio-alto y alto.	Medio-bajo y bajo
Infraestructura	Edificio escolar nuevo, amplio y apropiado. Escaso mantenimiento.	Edificio escolar relativamente nuevo, amplio y apropiado. Escaso para la matrí-cula. Muy bien man-tenido.	Edificio escolar antiguo, amplio y apropiado. Escaso mantenimiento.
Equipamiento	Escaso.	Suficiente y muy apropiado.	Escaso.

4. Análisis e interpretación

En esta parte expondremos los resultados obtenidos durante el desarrollo de la investigación. En primer lugar, efectuaremos una distinción inicial entre dos tipos de rituales escolares, a los que denominamos formales e informales:

Formales: constituyen rituales especialmente preestablecidos, siendo en este sentido actividades escolares incorporadas al tiempo y espacio escolar de un modo determinado, pudiendo por ello ser previstos. Encontramos entre estos las efemérides, los actos escolares, las formaciones y desfiles, entre otros.

Informales: son acontecimientos que transcurren en la vida cotidiana de la escuela, pero que por su eficacia simbólica se transforman en momentos especiales que realzan ciertos valores de la cultura.

En segundo término, precisaremos que no hemos trabajado una descripción exhaustiva de los rituales escolares, que consideramos innecesaria a los fines de nuestra investigación. En consecuencia, a continuación abordaremos seis Grupos de Rituales que hemos seleccionado, cuyo listado se incluye a continuación, dejando debidamente aclarado que estos constituyen solo un segmento del universo total que son los rituales escolares. Previamente debemos puntualizar que hablaremos de Grupos de Rituales, a los que definimos como: agrupaciones de rituales que poseen estructuras similares de significación y/o mecanismos comunes de realización. Dichos grupos son: Rituales del espacio y el tiempo, rituales de la domesticación de los cuerpos, rituales de las distinciones, rituales de los premios y castigos, rituales de la escritura, efemérides y actos escolares.

5. Discusión

En este apartado presentaremos nuestras conclusiones sobre la compleja telaraña que sustenta, anuda y consolida la compleja trama de lo escolar. Durante el desarrollo de esta investigación hemos tenido una clara evidencia sobre la existencia de un conjunto de rituales que signan la vida de la escuela. Los actores de la institución educativa los reproducen, los reelaboran, los reconstruyen y participan en ellos haciendo posible, de ese modo, que su eficacia simbólica se consolide. De esta manera, la institución educativa - en tanto lugar de condensación entre el individuo y la sociedad - crea lazos sociales, es decir socializa, reproduciendo en los nuevos miembros de una cultura los valores que esta sustenta. Los rituales son actividades esencialmente dramáticas cuya acción opera construyendo significados, poniendo de relieve, subrayando ciertos valores dominantes en una formación cultural dada.

Como ya hemos señalado, a lo largo de este estudio sistematizamos la recolección de datos, el análisis y la interpretación de seis grupos de rituales que atraviesan la vida escolar. Veamos ahora, en una síntesis hermenéutica, cuales son los rasgos destacables de cada uno de estos grupos.

Los rituales del espacio y del tiempo

Definidos como rituales que operan en la fragmentación y reticulación del espacio y el tiempo, manipulando las estructuras espacio-temporales de la acción, como modo de efectuar un control exhaustivo de los sujetos. Los rituales espaciales actúan desde diversas técnicas: la clausura, la zonificación, los emplazamientos funcionales, la distribución según rangos y el investimiento; pero en todo los casos esta tecnología disciplinaria tiende a constituir una topología del control, en la cual se delimitan territorios, se los segmenta y circunscribe, se les asignan valores y se define, en relación con los mismos, la inserción de los sujetos.

Los rituales del tiempo tienen como factor común, respecto a los espaciales, la reticulación. Esa acción sobre el tiempo se desarrolla, por medio de mecanismos específicos: el empleo del tiempo, la elaboración temporal del acto y la utilización exhaustiva. El ritual se desenvuelve a través del régimen de los horarios y del desdoblamiento de los tiempos en tiempos cada vez menores que implican actividades específicas, pero también lo hace pautando los tiempos de ejecución de las tareas, estimulando la mayor productividad en el menor tiempo y pone su acento en la homogeneización de los sujetos, sujetándolos a series temporales predeterminadas y ejerciendo de este modo el poder de control sobre los mismos.

Rituales de la domesticación de los cuerpos

Son rituales que ejercen su función sobre el disciplinamiento del cuerpo. En estos rituales el blanco del poder es el cuerpo y en este sentido cobra relevancia el concepto de lo social hecho cuerpo (Bourdieu). Domesticando el cuerpo se domestica al sujeto. Se apunta a establecer fuertes correlaciones entre el cuerpo y la disciplina, la tecnología de poder que circula a través de estos rituales opera desde el sistema de órdenes y señales, la correlación del cuerpo y del gesto, y la articulación cuerpo-objeto.

Rituales de las distinciones: **186**

Se tiende, mediante ellos a legitimar y subrayar las diferencias entre los alumnos. Estas diferencias no remiten a

reconocer la diversidad o valorizar la singularidad de los sujetos, sino a catalogar o rotular a los mismos en categorías que, generalmente, expresan diferencias de orden económico, social y cultural. Los rituales de las distinciones refuerzan mediante sistemas simbólicos de diversos tipos las diferencias existentes entre los alumnos y marcan la pertenencia de estos a uno u otro sector social.

Rituales de los premios y castigos

Su principal sentido es destacar la asociación entre las acciones de la vida escolar y el soporte principal de la modalidad disciplinaria (vigilar y castigar). Estos rituales ayudan a constituir un régimen de infrapenalidad en el cual se establecen las prácticas educativas aceptadas y por oposición las desviaciones de la norma. Infrapenalidad en tanto la escuela ha construido, a través de su historia, un exhaustivo conjunto de reglas que deben ser observadas, estas regulan el desempeño de los alumnos en la institución, pautando sus comportamientos con un nivel mayor de obsesividad que en la vida cotidiana. De ese modo, muchas acciones que para el conjunto de la vida social no constituyen faltas, detentan ese carácter al interior de la escuela. Este sistema de normas y desvíos posee su correlato en términos de premios y castigos, los que operan como recompensa o penalización de las acciones. Los alumnos no accionan en la escuela a partir de la significación que las propuestas educativas tienen para ellos, sino por sujeción a la norma, por el deseo de ser premiados o el miedo a ser castigados.

Rituales de la escritura

Este grupo de rituales presenta dispositivos similares a los de otros grupos pero se lo ha destacado por la especial atención que presta la educación de la modernidad a la escritura. En ellos, dos objetos culturales adquieren particular relevancia: los cuadernos y los libros de texto escolar. Estos objetos definen dos campos respecto a la escritura, el de lo que debe ser escrito y el de lo que debe ser leído.

En el primer caso, podemos observar como la alta ritualización de los cuadernos escolares ha determinado categorías de cuadernos y estilos de escritura, que subrayan la importancia del producto por sobre los procesos de construcción del conocimiento y erradican la aceptación del error, como punto de partida, para que esa construcción se opere. En cuanto a lo escrito, la estrategia del texto único denota una concepción acerca de la verdad, que es pensada unirreferencialmente y desde la versión oficial de la cultura dominante. El caso del texto único aborta toda posibilidad de lectura crítica, en el sentido freireano del término.

Efemérides y actos escolares

Entendemos que la identidad nacional es un arbitrario cultural presentado como orden autoevidente y natural. La nación es una comunidad imaginada y por lo tanto una representación arbitrariamente construida a partir de ciertos dispositivos como: la historia oficial, los mitos nacionales, los símbolos patrios, la religión, y los textos escolares.

Las efemérides ordenan el tiempo escolar de un modo diferente, lo hacen estableciendo una cronología que presenta, en un año, la totalidad de acontecimientos de la historia nacional y universal que la escuela asume deben ser destacados y recordados.

Los actos escolares son la puesta en escena, por excelencia, del conjunto de actividades que la escuela organiza como modo de promover la identidad nacional y los valores de la cultura. Valores estos que se definen desde los sectores hegemónicos de una formación social dada. En los dos casos es posible observar una alta asociación entre sus contenidos y el accionar de la modalidad disciplinaria. Las poesías, los himnos y las dramatizaciones memorizadas, las redacciones estereotipadas, las formaciones y saludos, son ejemplos de rituales que cabalgan velando más por los procesos que por los resultados. Al igual que en la domesticación de los cuerpos, en este caso se intenta vincular a los símbolos y fechas patrias con ciertas respuestas condicionadas, la nacionalidad penetra en los cuerpos, lo social se corporiza.

A modo de cierre

Adriana Puiggrós, al referirse al sentido de uno de sus estudios más interesantes sobre historia de la educación señala que el propósito principal de ese trabajo es colaborar en la recuperación de las alternativas pedagógicas " ... para que los educadores en cuyas manos se deshace la solemnidad educacional cada día, brotando algo nuevo, no se sientan tan solos en esta áspera actualidad argentina." (Puiggrós; 1994) Tal vez sin saberlo al iniciar este trabajo, nosotros comenzamos a transitar un camino parecido.

Notas

1. El presente trabajo es un resumen del Proyecto de Investigación desarrollado en la Facultad de Humanidades y Ciencias Sociales de la Universidad Nacional de Misiones (Argentina); que constituyó la Tesis elaborada por el autor, para acceder al Título de Magister en Educación, en la Facultad de Filosofia y Ciencias Humanas de la Universidad Católica Nuestra Señora de a Asunción (Paraguay).

2. DA MATTA, R. *CARNAVAÍS, MALANDROS E HEROÍS.* Editorial Zahar. Río de Janeiro, 1983. Pag. 24. La traducción es nuestra.

Bibliografía

ACHILLI, E., La práctica docente. Una interpretación desde los saberes del maestro, *Cuadernos de Antropología Social* . V 1. núm. 2. Facultad de Filosofía y Letras. U.B.A. Buenos Aires, 1988.

ALTHUSSER, L., *Ideología y aparatos ideológicos del Estado*, Nueva Visión. Buenos Aires, 1988.

ANDERSON, B., *Comunidades imaginadas*, Fondo de Cultura Económica. México, 1993.

APPLE, M.- GENTILI, P. y SILVA, T. T., *Cultura, política y currículo*, Losada. Buenos Aires, 1997.

BALL, S., *Foucault y la educación*, Disciplinas y saber, Morata. Madrid, 1993.

BOURDIEU, P. y PASSERÓN, J.C., *La reproducción*, Laia. Barcelona, 1981.

BOURDIEU, P., *Cosas dichas*, GEDISA. Buenos Aires, 1988.

BOURDIEU, P., *El sentido práctico*, Taurus. Madrid, 1991.

BOURDIEU, P., *Honor y gracia*, en: PITT-RIVERS, J. y PERISTIANY, J., Alianza. Madrid, 1992.

BRASLAVSKY, C., *La discriminación educativa en Argentina*, Miño y Dávila. Buenos Aires, 1994.

BROW, J. *Notes on Community, Hegemony and the Uses of the Past* (Traducción F. Jaume), 1990. (mimeograma).

BRIONES, G.*Métodos y técnicas de investigación para las ciencias sociales, Trillas.* México, 1992.

DA MATTA, R. *Carnavais, malandros e horois*, Zahar. Río de Janeiro, 1983.

DE ALBA, A. *Curriculum: crisis, mito y perspectivas*, Miño y Dávila. Buenos Aires, 1995.

DIAZ BARRIGA, A., *El currículo escolar*, REI/ IDEAS/ Aique. Buenos Aires, 1992.

DREYFUS, H. y RABINOW, P., *Michael Foucault: Más allá del estructuralismo y la hermeneútica*, Revista Mexicana de Sociología.

FAERMANN EIZIRIK, M. y COMERLATTO, D. *A escola invisível (Jogos de poder, saber e verdade)*, Editora da Universidade. UFRGS. Porto Alegre, 1995.

FILMUS, D. y FRIGERIO, G., *Educación, autoritarismo y democracia*, Miño y Dávila. Buenos Aires, 1988.

FOLLARI, R., *Práctica educativa y rol docente*, REI -Aique-IDEAS. Buenos Aires, 1992.

FOUCAULT, M., *Vigilar y castigar*, Siglo XXI. Buenos Aires, 1989.

FOUCAULT, M., *Diálogos sobre el poder*, Alianza Editorial. Buenos Aires, 1990.

FOUCAULT, M., *Microfísica del poder*, La Piqueta. Madrid, 1992.

FREIRE, P., *Pedagogía del oprimido*, Siglo XXI. México 1983.

FRIGERIO, G. (Compiladora) *De aquí y de allá. textos sobre la institución educativa y su dirección, Kapelusz.* Buenos Aires, 1995.

GARCÍA CANCLINI, N., *Las culturas populares en el capitalismo*, Casa de las Américas. La Habana, 1982.

GARCÍA CANCLINI, N., *Culturas híbridas. Estrategias para entrar y salir de la modernidad*, Grijalbo. México,

1990.

GEERTZ, C., *La interpretación de las culturas*, Gedisa. Barcelona, 1992.

GIDDENS, A. y otros., *La teoría social hoy*, Editorial Alianza. México, 1990.

GIROUX, H., *Teoría y resistencia en educación*, Siglo XXI. México, 1992.

GRAMSCI, A., *Antología*, Siglo XXI . México, 1988.

GUTIERREZ, A., *Pierre Boudieu. Las prácticas sociales*, Editorial Universitaria. Posadas, 1995.

LÉVI-STRAUSS, C., *Antropología estructural*, EUDEBA. Buenos Aires, 1977.

KEMMIS, S. en CARR, W. *Hacia una ciencia crítica de la educación, Laertes*. Barcelona, 1990.

LUKACS, G., *Historia y conciencia de clase*, Grijalbo. México, 1969.

McLAREN, P., *La escuela como una performance ritual*, Siglo XXI. México, 1995.

MARGULIS, M., *La cultura de la noche*, Espasa. Buenos Aires, 1994.

POPKIEWITZ, T., *Sociología política de las reformas educativas*, Morata. Madrid, 1994.

PUIGGRÓS, A., *Democracia y autoritarismo en la pedagogía argentina y latinoamericana*, Galerna. Buenos Aires, 1991.

PUIGGRÓS, A., *Sujetos, disciplina y curriculum*, Galerna. Buenos Aires, 1994.

PUIGGRÓS, A., *Volver a educar*, Ariel. Buenos Aires, 1995.

SAVIANI, D., *Pedagogía histórico-crítica. Primeras aproximaçoes, Cortez Editora*. San Pablo, 1991.

SAVIANI, D., *Educación: temas de actualidad, Libros del Quirquincho*. Buenos Aires, 1991.

TEDESCO, J. BRASLAVSKY, C. y CARCIOFI, R., *El proyecto educativo autoritario* (Argentina 1976-1982), *Miño y Dávila*. Buenos Aires, 1987.

TENTI FANFANI, E. *Los nuevos pobres,: efectos de la crisis en la sociedad argentina*, Losada-UNICEF. Buenos Aires, 1992.

TURNER, V., *La selva de los símbolos*, Siglo XXI. Madrid, 1980.

ZELMANOVICH, P. y otros, *Efemérides: entre el mito y la historia*, Paidós. Buenos Aires, 1996.

Acerca del Autoro

Pablo Daniel Vain

Professor in Universidad Nacional de Misiones. Director of the Master in University Teaching. General Secretary in the same University.
Profesor de Educación Física, Certificado de Estudios en Antropología

Social y Magister en Educación, realiza actualmente estudios de Doctorado en la Universidad de Málaga (España). Es Profesor Titular de Grado y Postgrado e Investigador de la Universidad Nacional de Misiones (Argentina), Director de la Maestría en Docencia Universitaria y Secretario General Académico de la misma Universidad. Ha desempeñado, además, diversos cargos en la gestión universitaria.

The World Wide Web address for the *Education Policy Analysis Archives* is http://epaa.asu.edu

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

190

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Japanese Higher Education Policy in Korea
During the Colonial Period (1910-1945)

Jeong-Kyu Lee
Korean Educational Development Institute
and
The Hongik University

Abstract
The purpose of this article is to examine the impact of Japanese nationalistic thought on the administrative systems and structures of colonial and modern higher education in Korea, as well as to analyze Japanese higher educational policy in Korea during the colonial period (1910-1945). It begins with an examination of Shinto, a syncretistic Japanese state religion and the ideological basis of national education. The author investigates Japanese educational policy and administration during the colonial period, including the establishment of a colonial imperial university in Korea. He also reviews the administrative systems and organizational structures in imperial and colonial universities. Both beneficial and negative impacts of the Japanese colonial education system on current Korean higher education conclude the analysis.

Shinto

Shinto was a spiritual foundation of the educational system of imperial Japan, as well as the national religion—or some would say cult. Throughout the history of Northeastern Asia, ancient Japan had close political, economic, and cultural relations with old Korea.Both Japanese*Nihongi* (Chronicles of Japan from the Earliest Times to AD 697) and *Kojiki* (Records of Ancient Matters) indicate numerous and multi-layered relationships between Korea and Japan. The earliest relations of Japan with the continent were mainly with Korea, particularly the Paekche Kingdom (18 BC-AD 660)[1], which was a cultural mediator between China and Japan (Hong, 1988; Longford, 1911; Maki, 1945). According to the records of Japanese Nihongi and Kojiki, Korea's two greatest early contributions to Japan were the transmission of Chinese writing and literature, and more importantly, Buddhism[2].The introduction of Buddhism had a significant effect on the development of Japanese culture and religion.A form of the northern branch of Buddhism (*Mahayana*) was transmitted to Japan via Tibet, China, and Korea (Aston, 1905, p. 359; Reader et al., 1993, p. 93). Indeed, Buddhism had a great impact on the development of Japanese culture as well as *Shinto*.

In the historical development of the Japanese religion and national thought, the origins of Shinto are highly controversial. Many eastern and western scholars (Aston, 1905; Holtom, 1938; Hong, 1988; Picken, 1994; Reischauer and Craig, 1973; Tsunoda et al., 1964)point out that Shinto cannot be separated from Buddhism, Confucianism, and

other continental influences[3].

In its earliest stage, Shinto was a primitive natural religion with elements of animism, natural worship, shamanism, ancestral reverence, agricultural rites, and purifications. Shinto later merged with Buddhism and Confucianism as *Ryobu* (Dual) Shinto[4], which contained religious and ethical components of a high order. Finally, the separation of Shinto from Buddhism was achieved, that is, *Kokka* (State) Shinto or *Jinja* (Shrine) Shinto as the state cult or religion (Aston, 1905; Bocking, 1996; Herbert, 1967; Holtom, 1938; Picken, 1994).Japanese ancestral worship is a combination of Shinto and Confucianism, what we call Shinto-Confucianism.

Twelve centuries later, *Shinto* was established under national and patriotic auspices and was subsequently adopted as Japan's national religion and ideology.[5] In 1870, Japanese Emperor Meiji issued a rescript defining the relation of Shinto to the state and the intention of the government concerning this matter. The Rescript states:

> We solemnly announce: The Heavenly Deities and the Great Ancestress...established the throne and made the succession sure. The line of Emperors...entered into possession thereof and transmitted the same. Religious ceremonies and government were of a single mind....Government and education must be made plain that the Great Way of faith in the kami [gods] may be propagated....(Holtom, trans., 1938, p. 55)

After the declaration of the Rescript, the Japanese government formulated the Three Principles of Instruction for the establishment of royal rule through a Shinto-centered indoctrination and decreed the Education Code for the foundation of modern educational systems. On April 28, 1872, the Education Code was proclaimed: (1) compliance with the spirit of reverence for *Kami* (Gods) and love of country; (2) clarification of 'the principle of Heaven and the Way of man'; and (3) exalting the Emperor and obeying the Imperial Court (Tsunetsugu, 1964, p. 206). The 1872 Education Code of Japan emulated the uniform and centralized system of France initiated by Napoleon III in 1854 (Anderson, 1975, p. 21).

Furthermore, the Japanese government attempted to set up national morals within the schools based on the Shinto-Confucian Imperial Rescript on Education[6] that was promulgated on October 30, 1890 (Anderson, 1959, p. 13; Beauchang & Vardaman, 1994, pp. 4-5; Holtom, 1938, p. 71; Horio, 1988). The Rescript stressed the Shinto ideology of royal worship mixed with Confucian ethical concepts and practices such as loyalty, filial piety, benevolence, ancestor worship, learning, and harmonious human relationships. Shinto appealed to Japanese cultural nationalists because it combined ethical codes of virtue and honor with an even more exalted ethic of duty to the state, and to the divinely inspired head-of-State in particular.

Therefore, Shinto ideology and Confucian concepts were two main pillars of Japanese imperial education. The Meiji Rescript, as a Holy Writ or a national moral prop of the Japanese people, was reinterpreted several times in maintaining the rising militaristic and ultranationalistic ideology. Its philosophy extended to the educational systems of Japan. State-Shinto or National-Shinto dictated an administrative structure in government as well as in higher education that enforced a strict stratification system, centralized governance, and intellectual conformity. Certainly these features were reinforced further, even ossified with the Japanese occupation and colonialization of Korea from 1910 to 1945. Japanese imperialists set up the ruling policy that aimed to let Koreans assume the personalities of loyal citizens of her imperialism. To fulfill their political scheme, the Japanese nationalists imposed Shinto-Confucianism on Korea and attempted to design a new educational system and an administrative structure suitable for the execution of their colonial policy. Therefore, higher education was an essential tool in accomplishing the Shinto-Confucian ideologies during Japanese colonization.

Japanese Educational Policy and Administration in Colonial Higher Education

After the 1895 Shimonoseki Treaty, Japan introduced western-style institutions and reforms Kora including the elimination of such social practices as class discrimination. However, these reforms were met with hostility from a broad cross-section of Koreans who felt their traditional Confucian and shamanistic beliefs threatened by the social-leveling tendencies of western-style democracy. Having won the Russo-Japanese War in 1905, Japan moved immediately to establish a protectorate over Korea, called the 1905 Protectorate Treaty (Kibaek Lee, 1984, p. 309). After the treaty was signed, the Choson government nearly lost its national right to govern. During the 'Protectorate' period (1905-1910), the Japanese educational policy was chiefly the preparatory operation for colonization through the promulgation and practice of various educational ordinances and regulations. For instance, the Private School Ordinance (*Sarip-hak-kyo-ryeong*), which was promulgated in 1908, was a means of placing under Japanese control and suppression all the private schools administered by Christian missionaries and patriotic Korean leaders (KNCU, 1960, p. 15).

In 1911, the Japanese colonial government proclaimed the Educational Ordinance[7] in accordance with the Imperial Rescript (Cheong, 1985, p. 283; Keenlyesids and Thomas, 1937, p. 100; Sung-hwa Lee, 1958, pp. 83-84; Nam, 1962, p. 38; The Government-General of Choson, 1935, p. 167; Yu, 1992, p. 126).The Educational Ordinance appeared as follows:

193

Be filial to your parents, affectionate to your brothers and sisters; as husbands and wives be harmonious, as friends true; bear yourselves in modesty and moderation; extend your benevolence to all; pursue learning and cultivate the arts, and thus develop your intellectual faculties and perfect your morality. Furthermore, be solicitous of the commonwealth and of the public interest; should emergency arise, offer yourselves courageously to serve the State.(Keenlyeside and Thomas, 1937, p. 100).

Based on the above Ordinance, the Japanese colonial administration urged elementary, secondary, and vocational education, including medical, foreign language, and teacher education. The Educational Ordinance of 1911 allowed higher educational institutions, such as Christian missionary colleges, to lose their college statuses and be downgraded to non-degree granting schools. It was not until the promulgation of a new Educational Ordinance on February 4, 1922 that previous higher educational institutions were accredited once again.

The ordinance was a strategy by the Japanese to force the Korean people to become compliant to Japanese imperialism, to undermine the nationalism of Koreans, and ultimately to transform the people into loyal Japanese citizens. After issuing a new Educational Ordinance on February 4, 1922, several Christian missionary schools and one Korean private collegiate school that had lost their college statuses were upgraded as college institutions.The major difference between the old (1911) and the new ordinances (1922) was that the latter abolished a dual discriminative system and applied the Japanese educational system throughout Korea.

At the same time, patriotic Korean leaders promoted an educational movement to implement their own private colleges or universities (Lee, 1965, p. 241). To offset this trend, the Japanese administration opened Keijo Imperial University (now evolved into Seoul National University) in 1924, under the Ordinance of University, and based on the Meiji Rescript (The Government-General of Choson, 1935, p. 486). This was to be the first modern university in Korea, which included the departments of law and literature, and medicine. Although the Japanese established a new national-level university in Seoul, most Koreans, nationalists and conservative Confucians, did not enroll their sons and daughters in the new imperial university. Instead, many patriotic intellectuals who were eager to encourage nationalism opened several private schools. These open, night, and labor schools were designed for Koreans to enhance national spirit.

The Japanese colonial government claimed that Keijo Imperial University in Seoul was almost the same as Imperial universities in Japan in terms of quality (The Government-General of Choson, 1935, p. 486), but the university was not a scientific research institute like the Japanese imperial universities. In truth, Tokyo Imperial University, as a scientific research university, was organized into four departments: law, science, literature, and medicine. Despite the fact that Keijo University was a prototype of a Japanese imperial university, it became a model for successive modern Korean universities.

Regarding educational structure and systems, an Educational Bureau under the Internal Affairs Department in the Government-General of Choson became a top organ of educational administration after Japanese annexation. The Educational Bureau was composed of an educational section, an editorial section, a religious section, and a school inspectorate. In the provinces, educational sections formed part of the Department of Internal Affairs and had a staff of school-inspectors (The Government-General of Choson, 1921, p. 75). The chief of the Educational Bureau was controlled and supervised by the Director of Internal Affairs, who was in charge of the entire educational system of Korea (Cynn, 1920, p. 100). Educational administration under Japanese rule was highly centralized in the Internal Affairs Department and in the Educational Bureau, and was directed and supervised by these offices due to their coercive power within the organizational hierarchy. The Educational Bureau under the Internal Affairs Department had responsibility for most aspects of the whole school system, including missions and aims, scholastic terms, curricula, qualifications of teaching staff, management of personnel, fiscal review, allotment of funds, and inspection of educational facilities.

Administrative control of educational affairs such as policy-making, establishment of schools, compilation and censorship of textbooks, granting of teacher certificates, hiring and assigning of teaching staff, formation of the educational budgets and approvals, and scholarship administration were exercised on the authority of the Government-General of Choson (The Government–General of Choson, 1921, 1935).

Top policy of the Japanese Emperor was issued in Imperial Ordinances prepared by the governor of the Government-General of Choson. Policy change was usually initiated in the form of directives and instructions by the department and bureaus under the Government-General (Anderson, 1959, p. 75). The administrators of these offices stressed authoritative hierarchical orders that were followed without questions by the subordinates of the organizational systems.

During the Japanese occupation, the highly centralized system of educational administration based on Imperial Ordinances was used to reinforce centralized governance and intellectual conformity, as well as to eliminate Korean nationalism, independence, and cultural identity. The Japanese educational system and structure was a means to edify the Korean people in accordance with the Meiji Rescript on Education. Thus, the colonial educational system and structure were tools to achieve Japanese political schemes, denationalization and assimilation.

Administrative System and Organizational Structure in Imperial and Colonial Universities

Under the imperial Japanese rule, there were nine imperial universities. Seven of these were in Japan: Tokyo (1886), Kyoto (1897), Kyushu in Fukuoka (1903), Hokkaido in Sapporo (1903), Tohoku in Sendai (1909), Osaka (1931), and Nagoya (1931). Two were located in the colonies: Keijo in Korea and Taihoku in Taiwan (Anderson, 1959, p. 126). The governing system and the organizational structure of Keijo Imperial University were copied directly from Japanese Imperial Universities, which were patterned after several Western countries' academic models and institutions, particularly Germany (Altbach, 1989; Anderson, 1959; Cummings, 1990). Many ideas and models of higher education were taken from Western countries, including French administrative organizations and bureaucratic coordination systems, Pestalozzi's developmental educational system, Herbartian moral centered-pedagogy, German university models and structures for academia, Anglo-American ideas of utilitarian education, American liberal arts philosophy and American pragmatism, especially John Dewey's educational philosophy (Altbach, 1989; Cummings, 1990, p. 73; Cummings, Amano, & Kitamura, 1979; Nakayama, 1989, pp. 31-48).

Chinese educational ideas based on Confucianism and Chinese classics also had a great impact on Japanese education. Indeed, after adopting many Western ideas of higher education, the Japanese incorporated them into the Shinto-Confucian tradition. Shigeru Nakayama (1989), a Japanese historian, asserts that "the first example of the window-shopping mode occurred in the late nineteenth century, whereas the involvement mode is best illustrated in the post-World War II Occupation period, in which reforms based on the American system were carried out" (pp. 31-32).

Japanese Imperial higher education adopted the centralized system of France, as well as a system of rank structure modeled on the German approach (Anderson, 1975, p. 21; Cummings, 1990, p. 113). Keijo Imperial University as a colonial institute was also shaped by a highly centralized organizational structure. The entire academic structure was set up in accordance with the Japanese prototype. Accordingly, the curriculum of Keijo Imperial University was almost identical to the Japanese imperial universities and the majority of the academic staff and students were Japanese (The Government-General of Choson, 1935). Furthermore, educational administrators and faculty members used the Japanese language for higher education, including teaching and learning, textbooks, and communicating with faculty members. Not only did the Japanese colonial administrators manage academic affairs and finance, but they also supervised closely all faculty members from the president to the administrative and teaching staff (The Government-General of Choson, 1935, p. 486). The Japanese administrators appointed all faculty regarding their working positions and functions, and controlled students' activities and academic freedom (Ibid).

In terms of educational administration, the administrative system and structure of Keijo University was almost the same as that of the Japanese university. Like the metropolitan imperial universities, Keijo University was hierarchical in organization and had an authoritative system of rank structure. The university administrators and the colonial authorities imposed strict rules, and hierarchical authority through royal rescripts, ordinances, policies, and directives. As Cummings (1990) points out, the system comprised a linear rank structure in which the head of the chair exerted absolute authority. Further, the academic ranks of professor, assistant professor, instructor, assistant, and vice-assistant taught and assisted in each field. In the selection of new faculty members, the most important criterion was age. The age rank structure based on Confucian ethical and social values solidified authoritarian leadership of top and middle line senior administrators. Accordingly, the open-rank system, which depended on cooperation and more objective evaluations, was not practiced.

In this manner, the organizational structure of Keijo Imperial University was maintained in a highly centralized formal system based on Shinto-Confucian values and norms. In addition, the Meiji Rescript was a blueprint of the Shinto-Confucian educational plan and a seed of institutional culture in colonial higher education.

Under Japanese colonial rule, Koreans were discriminated against either in institutional programs or training. The Japanese imperial administration offered higher educational opportunities to the Japanese people. Few Koreans could access elite (Lee, 1984, pp. 367-68). Actually, Japanese administrators under restrictive administration and curriculum policies provided Koreans with few chances to enter higher educational institutes and did not educate them in advanced engineering and scientific courses. In 1925, at the college level, the proportion of Korean enrollment was no more than one-twenty-sixth of Japanese and at the university level over one-one hundredth (Lee, 1984, p. 367). Japanese used two educational systems to discriminate between Japanese and Koreans: one was an educational system for persons using Japanese, and the other for persons using Korean. As Jin-Eun Kim (1988) points out, the Japanese were allowed to operate within a separate privileged system, while the Koreans were subject to limitations in secondary and higher education.

At that time, university admission of Korean people was strictly limited, and only very few Koreans, who were by and large the offspring of pro-Japanese persons or rich people, attended Keijo University. Many scions of pro-Japanese and rich people enrolled at Japanese vocational or teachers' schools. Sungho Lee (1989) points that "the total enrollment of the Keijo Imperial University in 1934 in ten years since its establishment was 930, of which the Korean fraction was only 32 percent" (p. 95).

Specifically, in 1939, there were only 0.27 Korean students in colleges and teachers' training seminaries for every 1,000 Koreans of the general population, and 7.20 Japanese students for every 1,000 Japanese in Korea. There were 0.0093 Korean students enrolled in university for every 1,000 Koreans, while 1.06 Japanese university students per 1,000 of Japanese population in Korea (Grajdanzev, 1944, p. 264; Sungho Lee, 1989, p. 94; UNESCO, 1954, p. 24).

The higher educational schools under Japanese colonial rule were viewed by the nationalistic Koreans as training institutes that cultivated pro-Japanese agents serving the Japanese imperialists. In fact, as Byung Hun Nam (1962) mentions, the primary motives of the establishment of this university were to offer higher education for the Japanese in Korea, to forestall growing Korean nationalism, and to indoctrinate the Korean elite as pro-Japanese. Indeed, some of the Koreans who studied at Keijo University faithfully served Japanese imperialists as puppets or collaborators during the Japanese colonial period (Chang, 1992; Seo, 1989). For instance, among 804 Korean graduates, 228 persons served at Japanese governmental and public offices (Chang, 1992, p. 392).

In particular, during World War II (1937-1945), the Japanese regime announced three educational principles of its administration. These included profound understanding of the national mission, strengthening Japanese and Korean unity, and dedication to labor for the realization of national goals. Japanese militarism reached its peak following the establishment of the puppet government of *Manchukuk*. The Japanese colonial administration demanded that the Korean people, including Western missionary teachers and students, should pay homage to Shinto shrines (Palmer, 1977, pp. 139-40). They forcibly demanded that the Koreans should use the Japanese language, instruct all classes in Japanese, and change their traditional family names to reflect Japanese styles (Meade, 1951, p. 213).

From this, it can be concluded that the purposes of the Japanese colonial education were denationalization, vocationalization, discrimination, and assimilation, according to Han-Young Rim's (1952) analysis. Especially at the higher level, the ultimate goal of university education in Korea was to foster the pro-Japanese elite as faithful Japanese puppets. Furthermore, after the liberation in 1945, these Japanese agents ironically became the privileged class leading to a new Korean society (Chang, 1992; Cheong, 1985; Choi, 1990; Im, 1991; Lee, 1985; Lee, 1997; Seo, 1989). For example, during the 12 years Syngman Lee's administration (1948-1960), 83 percent of 115 Cabinet ministers were Japanese agents or collaborators under Japanese colonial rule (Seo, 1989, p. 452).

On the contrary, many patriotic or nationalistic Korean people participated in the army for national independence, and attended the native private schools, or Christian missionary institutes instead of Japanese institutes. More than half of the Korean students attended private Korean colleges or collegiate schools, and many of them journeyed abroad[8] to access higher education (Lee, 1984, p. 368). In fact, many Confucian learned men were actually reluctant to accept Western education, resulting in their adherence to the Confucian educational tradition at village schools[9].

The Impact of the Japanese Colonial Education System on Current Korean Higher Education

During the Japanese colonial period (1910-1945), Japanese imperialists designed the educational system and administrative structure to reflect a Shinto-centered philosophy. This was used as a tool to aid the assimilation of Koreans to a more Japanese point of view and subverted the Korean national spirit. Shinto ideology was integrated into colonial higher education through emphasis on the worship of Shinto shrines as well as Shinto-Confucian concepts in the college curriculum. With the enforcement of the cultural assimilation policy and practice, Japanese colonizers used Shinto ideology as a means of strict disciplinary action against Koreans, eliminating freedom of speech, clamping down on colleges or universities, and eradicating Korean nationalism. The resulting tensions among Korean nationalism, independence, and democracy were at the heart of Korean educational development in the twentieth century.

In addition, the Japanese authorities offered higher education opportunities to some pro-Japanese Koreans to train as an elite group who could support the pro-Japanese militarism. Despite such an undesirable policy, the heritage of Japanese colonialism shaped the nature of the modern Korean universities and left both positive and negative outcomes within Korean higher education.

The positive effects were that the Japanese colonial government established several collegiate institutions including a university, endorsed public education for many Koreans regardless of social status and gender, introduced Western technical and professional training through common higher or collegiate level institutes, and transferred preferred administrative systems and practices. The administrative system and structure became models for modern Korean higher education. Many Korean intellectuals who had studied at the colonial or Japanese imperial universities played an important role in the foundation of contemporary Korean higher education (Banminjokmoonjeyeonkuso, 1993; Chang, 1992; Cheong, 1985; Choi, 1990; Im, 1991)[10].

Several negative results can also be noted. Firstly, Japanese colonial authorities regarded higher education as a tool to foster pro-Japanese elite agents who were able to practice Japanese colonial policy and Japanese imperialism based on Shinto-Confucianism. Secondly, the Japanese abolished the Confucian National Academy which had preserved the Korean academic tradition. Thirdly, Korean tertiary institutes under the Japanese colonial period lost opportunities to introduce Western models which may have been well suited for Koreans' needs. Finally, some Korean alumni of the Keijo Imperial University became pro-Japanese collaborators, resulting in unfair or discriminatory practices for Korean educators (Banminjokmoonjeyeonkuso, 1993; Chang, 1992; Choi, 1990; Im, 1991; Lee, 1985)[11].

In terms of educational administration, a closed organizational system--rigid and authoritative leadership, a hierarchical centralized formal sturcture, closed communication networks, and administrator-centered education--has formed the organizational system and culture in contemporary Korean higher education. Moreover, several Western education systems, for example, "window shopping modes," adopted by the Japanese are the typical types of

BEST COPY AVAILABLE

administrative systems in current Korean higher education. For instance, a centralized system and a linear rank structure are the backbones of the organizational systems in the Ministry of Education and higher education institutions.

In particular, the Meiji Rescript on Education promulgated by the Japanese Emperor Meiji in 1890 was a matrix of the Chart of National Education[12] promulgated for the recovery of national spirit and educational reform by the Park Administration in 1968. The Chart was a guiding principle in Korean education from the 1968 until the early 1980s. In addition, Keijo Imperial University established by the Japanese Colonial Administration in 1924, was a precursor of the present Seoul National University, and has produced a large number of bureaucrats and talents as leading individuals who play important roles in the present Korean society.

From all of this, it can be concluded that the story of Japan's influence in Korea and the historical connection between these two traditional rivals is far more complex and nuanced than the paper would suggest. That history is replete with rich and telling ironies. The importance of Shinto, a syncretistic Japanese state religion borrowing elements not only of Chinese Confucianism but also of Korean Buddhism and Shamanism, is a good case in point. In addition, Japan undertook to introduce Western-style institutions and reforms, including the elimination of such social practices as class discrimination. Furthermore, State-Shinto dictated an administrative structure in higher education as well as in government that enforced a strict stratification system and intellectual conformity, and even ossified the Japanese colonialization of Korea 1910-1945.

The Japanese reforms based on the ideology of State-Shinto met with general hostility from a broad cross-section of Koreans, who felt their traditional Confucian and national beliefs being threatened by the social-leveling tendencies of western-style democracy. After 1910, Japanese colonizers took a harder line against Koreans, eliminating freedom of speech, the press, and association and clamping down on the universities. This caused public resistance, but around the issue of independence, not the restoration of western-style freedoms. The tensions between Korean nationalism, independence, and democracy are at the heart of the story of Korean educational development and yet remain largely unexplored. Clearly, the heritage of Japanese colonialism has contributed to the shaping of the administrative systems of the contemporary Korean universities and also has positively and negatively affected overall current Korean higher education.

Notes

[1] In the history of Korea, Paekche Kingdom (18 BC-AD 660), as one of Three Kingdoms, was located in the southwest of the Korean peninsula. Three Kingdoms were Koguryo (37 BC-AD 668) in the north; Silla (57 BC-AD 935) in the southeast; and Paekche. Silla unified the Korean peninsula. The next epoch was Koryo Kingdom (918-1392), and the last Korean kingdom was Choson (1392-1910).

[2] Like Nihongi's records (Vol. I, pp. 262-63), Kojiki also left Wangin (Wani)'s contribution in AD 285 (Aston notes that the year corresponds to AD 405). Kojiki describes that the King of Packche presented a man named Wani-kishi, and by this man he presented the Confucian Analects in ten volumes and the Thousand Character Essay in one volume (tr. Chamberlain, p. 306). In AD 552, the Nihongi records that the King of Paekche in Korea sent an embassy to Japan with a present to the Mikado of an image of Shaka Buddha in gold and copper, banners, umbrellas, and a number of volumes of the Buddhist Sutras (tr. W. A. Aston, pp. 59-60).

[3] Wontack Hong (1988), a Korean historian, claims that "The dominant religion in Korea prior to the introduction of Buddhism and Confucianism was Shamanism. This Shamanism seems to have been brought to Japan by those who migrated from Korea" (pp. 138-39). Ryusaku Tsunoda and William T. de Bary (1964) also claim that "Shinto was not an indigenous religion...Shamanistic and animistic practices similar to these of Shinto have also been found through northeast Asia, especially in Korea" (p. 21). In addition, Edwin O. Reischauer and Albert M. Craig (1973) assert that "[m]embers of the priestly class who performed the various rites...probably represented the Japanese variant of the shamans of Korea and Northeast Asia" (p. 473). Lastly, W. G. Aston (1905), a translator of Nihongi, insists that in prehistoric Shinto, there are definite traces of a Korean element in Shinto A Kara no Kami (God of Korea) was worshipped in the Imperial Palace (p. 1). Stuart D. B. Picken (1994) mentions that "Shinto has been described as the source of Japan's creative spirit on the one hand, and as an incorrigible source of militaristic nationalism on the other" (p. 4).

[4] Ryobu Shinto means "Two-sided" or "Dual Shinto." A Popular Dictionary of Shinto (Bocking, 1996) notes, "An interpretation of Kami (Gods) beliefs and practices developed in the Kamakura period (1185-1333) and maintained by the Shingon School of esoteric Buddhism. A derivative theory that reversed the status of kami and Buddhas was proposed by Kanetomo Yoshida (1435-1511)" (p. 145).

[5] After the Meiji Restoration of 1868, the affairs of both Shinto and Buddhism were placed under the same set of official regulations on April 21, 1872 (Holtom, 1938, p. 59). However, in February 1873, the Japanese government proclaimed officially that it would protect the freedom of Shinto and Buddhism and that it encouraged each of them to grow (Herbert, 1967, p. 51). Brian Bocking (1996) notes: "'State Shinto,' 'National Shinto,' or 'Shrine Shinto' was a concept defined retrospectively and applied by the occupation authorities in the Shinto Directive of 1945 to the post-Meiji religious system in Japan. In the Directive, State Shinto is defined as 'that branch of Shinto (Kokka Shinto or

Jinja Shinto) which by official acts of the Japanese Government has been differentiated from the religion of sect Shinto (*Shuha* Shinto) and has been classified a non-religious cult commonly known as State Shinto, National Shinto, or Shrine Shinto'" (pp. 100-01).

[6] The Japanese Emperor Meiji's Rescript on Education notes:

> Know ye, Our Subjects:
>
> Our Imperial Ancestors have founded our Empire on a basis broad and everlasting, and have deeply and firmly implanted virtue; Our Subjects ever united in loyalty and filial piety have from generation to generation illustrated the beauty thereof...Ye, Our Subjects, be filial to your parents, affectionate to your brothers and sisters; as husbands and wives be harmonious, as friends true; bear yourselves in modesty and moderation; extend your benevolence to all; pursue learning and cultivate arts, and thereby develop intellectual faculties and perfect moral powers; always respect the Constitution and observe the laws; should emergency arise, offer yourselves courageously to the State; and thus guard and maintain the prosperity of Our Imperial Throne coeval with heaven and earth....(Sansom, trans.,1950, p. 464)

[7] With the Meiji Rescript, the Educational Ordinance was a fundamental frame for governing colonial education in Korea until August 15, 1945, although the Japanese colonial administration revised and enacted several educational ordinances in 1922, 1938, and 1943 (Cheong, 1985; Jin-Eun Kim, 1988; Nam, 1962; Yu, 1992).

[8] In 1931, 3,639 Korean students were enrolled in Japanese tertiary institutions, whereas as many as 493 Koreans were studying in the United States (Lee, 1984, p. 368).

[9] In the history of Korea, Confucian education traditionally maintained two streams from the Three Kingdoms period to the early twentieth century. One stream was of national institutions, and the other stream was of civil or village schools. The national Confucian institute, Seongkyunkwan, was compulsorily abolished by the Japanese imperialists in the early twentieth century, but many Confucian civil or village schools actually existed in the provincial areas during the Japanese colonial period.

[10] When the United States Military Government organized Korean Committee on Education in September 1945 in order to build a new Korean education, the majority of committee members were pro-Japanese collaborators who studied in Japanese imperial universities during the Japanese occupation (Banminjokmo-onjeyeonkuso, 1993; Cheong, 1985, pp. 85-88; Im, 1991). Furthermore, many graduates of the colonial and imperial universities became faculty members of the new university when Keijo University evolved into the Seoul National University in 1946 (Choi, 1990, p. 51).

[11] Many Korean alumni became Japanese governmental or public officers and suffered the Korean people (Chang, 1992; Lee, 1985). For instance, H. N. Lee, an alumnus of Korean Imperial University, was a county magistrate who drafted young Koreans for the Japanese Pacific War under the rule of Japanese imperialism, but he became a professor and president at a university in Seoul under the contemporary Korean government (Chang, 1992, p. 348). B. D. Jeon, as a public officer in Kyungki province, suppressed many patriotic Korean nationalists in the Japanese colonial period (Chang, 1992, p. 394).

[12] The Chart states:

We have been born into this land, charged with the historic mission of regenerating the nation...With the sincere mind and strong body improving ourselves in learning and arts, developing the innate faculty of each...we will cultivate our creative power and pioneer spirit. We will give the foremost consideration to public good and order, set a value of efficiency and quality, and inheriting the tradition or mutual assistance rooted in love, respect and faithfulness, will promote the spirit of fair and warm cooperation... The love of the country and fellow countrymen together with the firm belief....we pledge ourselves to make new history with untiring effort and collective wisdom of the whole nation. (Ministry of Education, 1976, p. 3)

References

Altbach, Philip G. (1989). Twisted Roots: The Western Impact on Asian Higher Education. *Higher Education, 18*: 9-29.

Anderson, Ronald S. (1959). *Japan: Three Epochs of Modern Education*. Washington D. C.: United States Government Printing Office of Education.

Anderson, Ronald S. (1975). *Education in Japan*. Washington D. C.: U. S. Government Printing Office.

Aston, W. G. (1905). *Shinto (The Way of the Gods)*. New York: Longmans, Green, and Co.

Banminjokmoonjeyeonkuso (1993). *Chinilpa 99 In (99 Pro-Japanese Koreans) 1, 2, 3*. Seoul, Korea: Dolbege.

Beauchang, Edward R., & Vardaman, James M. Jr. (Eds.). (1994). *Japanese Education Since 1945: A Documentary Study*. Armonk, New York: M. E. Sharpe Inc.

Bocking, Brian. (1996). *A Popular Dictionary of Shinto*. Surrey: Curzon Press.

Chang, S. Y. (1992). *Ilje-ui Kyungsung-Jekuk Daehak Seolipkwa Woonyoug (The Establishment and Management of Keijo Imperial University)*, *Hankuk-Dokrip-Woondongsa-Yeonku, Vol. 6*, pp. 347-408.

Cheong, Jae-cheol. (1985*). Ilje-ui Dae-hankuk-Sikminji-Kyoyuk-Jeong-cheksa (A History of Colonial Educational Policy in Korea under Japanese Imperialism)*. Seoul, Korea: Iljisa.

Choi, Hojin. (1990, August). *Kyungseongdae Jekeon-gwa Kukdaean-padong-ui Oajung-eseo (Rebuilt of Kyungseong University under the Vortex of a Kukdaean Wave)*. *Shin-Dong-A, 8*, 248-60.

Cummings, William K. (1990). *The Changing Academic Marketplace and University Reform in Japan*. New York: Garland Publishing.

Cummings, William K., Amano, Ikuo & Kitamura, Kazuyuki. (1979). *Changes in the Japanese University: A Comparative Perspective*. New York, New York: Praeger Publishers.

Cynn, Heung-wo. (1920). *The Rebirth of Korea*. New York: Abingdon Press.

Grajdanzev, Andrew J. (1944). *Modern Korea*. New York: The John Day Company.

Herbert, Jean. (1967). *Shinto: At the Fountain-Head of Japan*. New York: Stein and Day.

Holtom, D. C. (1938). *The National Faith of Japan*. New York: E. P. Dutton & Company.

Hong, Wontack. (1988). *Relationship Between Korea and Japan in Early Period: Paekche and Yamato Wa*. Seoul, Korea: Ilsisa.

Horio, Teruhisa. (1988). *Educational Thought and Ideology in Modern Japan: State Authority and Intellectual Freedom*. (Steven Platzer Ed., & Trans.). Tokyo, Japan: University of Tokyo Press.

Im, Chong-Kuk (1991). *Silok Chinilpa (Authentic Stories of Pro-Japanese Koreans)*. Banminjokmoonjeyeonkuso (Ed.), Seoul, Korea: Dolbege.

Keenlyesids, Hugh & Thomas, A. F. (1937). *History of Japanese Education*. Tokyo, Japan: The Hoku Sei-do Shoten.

Kim, Jin-Eun. (1988). South Korea. In George Thomas Kurian (Ed.), *World Education Encyclopedia, Vol. II*. New York: Facts on File Publications.

Kojiki (Records of Ancient Matters). (1973). (Basil Hall Chamberlain, Trans.). Tokyo, Japan: The Asiatic Society of Japan.

Lee, Chong Sik. (1965). *The Politics of Korean Nationalism*. Berkeley: University of California Press.

Lee, Jeong-Kyu. (1997). *A Study of the Development of Contemporary Korean Higher Education*, Unpublished Ph. D. dissertation, The University of Texas at Austin.

Lee, Kibaek. (1984). *A New History of Korea*. (Edward W. Wagner & Edward J. Shultz, Trans.). Cambridge: Harvard-Yenching Institute.

Lee, Kidong (1985). *Iljeha-eui Hankukin Kwanrideul (Korean Officers under the Japanese Imperialism)*. Shin-dong-A, 3, 472-75.

Lee, Sungho. (1989). The Emergence of the Modern University in Korea. *Higher Education, 18*, 87-116.

199

Lee, Sung-hwa. (1958). *The Social and Political Factors Affecting Korean Education (1885-1950)*. Unpublished Doctoral dissertation, University of Pittsburgh.

Longford, Joseph H. (1911). *The Story of Korea*. London: Adelphi Terrace.

Maki, John M. (1945). *Japanese Militarism*. New York: Alfred A. Knopf.

Meade, E. Grant. (1951). *American Military Government In Korea*. New York: Columbia University Press.

Minisry of Education (1976). *Education in Korea*. Seoul, Korea.

Nam, Byung Hun. (1962). *Educational Reorganization in South Korea under the United States Army Military Government, 1945-1948*. Unpublished Doctoral dissertation, University of Pittsburgh.

Nakayama, Shigeru. (1989). Independence and Choice: Western Impacts on Japanese Higher Education. *Higher Education, 18*, 31-48.

Nihongi (Chronicles of Japan from the Earliest Times to AD 697). (1896). *Vols. I & II*. (W. G. Aston, Trans.). London: The Japan Society.

Palmer, Spencer J. (1977). Korean Christians and the Shinto Shrine Issue. In C. I. Eugine Kim & Doretha E. Mortimore (Eds.), *Korea's Response to Japan: The Colonial Period 1910-1945*. Western Michigan University: The Center for Korean Studies.

Picken, Stuart D. B. (1994). *Essentials of Shinto: An Analytical Guide to Principal Teachings*. Westport, Connecticut: Greenwood Press.

Reader, Ian, Andreasen, Esben, & Stefansson, Finn. (1993). *Japanese Religions: Past and Present*. Sandgate, Folkestone, Kant, England: Japan Library.

Reischaure, Edwin O., & Craig, Albert M. (1973). *Japan: Tradition & Transformation*, Boston: Houghton Mifflin.

Rim, Han-Young. (1952). *Development of Higher Education in Korea during the Japanese Occupation (1910-1945)*. Unpublished Ed. D. dissertation. Teachers College, Columbia University.

Sansom, G. B. (1950). *The Western World and Japan*. New York: Alfred A. Knopf.

Seo, Byung-Uk. (1989, October). *Minjok-Jeongki Salyeoja-hammida (We Should Revive A National Spirit)*. Shin-Dong-A, 10, 444-455.

The Government-General of Choson. (1921). *Annual Report on Reforms and Progress in Chosen (1918-21)*. Keijo.

The Government-General of Choson. (1935). *A History of 25 Year Administration*. Keijo.

The Korean National Commission for UNESCO [KNCU]. (1960). *UNESCO Korean Survey*. Seoul, Korea: The Dong-A Publishing Co., Ltd.

Tsunetsugu, Muraoka. (1964). *Studies in Shinto Thought*. (Delmer M. Brown & James T. Araki, Trans.). Japan: Ministry of Education.

Tsunoda, Ryusaku & de Bary, William T., & Keene, Donald. (1964). *Sources of Japanese Tradition, Volume I*. New York: Columbia University Press.

Underwood, Horace H. (1926). *Modern Education in Korea*. New York, New York: International Press.

UNESCO. (1954). *Rebuilding Education in the Republic of Korea*. Frankfurt am Main: Johannes Weisbecker.

Yu, Bongho. (1992). *Hankuk Kyoyuk Gwajeongsa Yunku (A Study of the History of Curricula in Korea)*. Seoul, Korea: Kyohak-yunkusa.

About the Author

Jeong-Kyu Lee
Associate Research Fellow
The Division of School Education Research
Korean Educational Development Institute (KEDI)

Joint Professor of Higher Education Administration
Hongik University
Seoul, Korea.

Phone: (02) 3460 – 0382; Fax : (02) 3460-0117
E-mail: jeongkyuk@hotmail.com

The author received a Doctor of Philosophy degree in Higher Education Administration at the Graduate School of The University of Texas at Austin in the United States of America. He is Associate Research Fellow in the Division of Educational Policy Research at the Korean Educational Development Institute, a Korean government-funded research institute, in Seoul, and Joint Professor at the Graduate School of Educational Management at Hongik University in Seoul, South Korea. His work has been published in a variety of national and international journals. His research interests are organizational culture and leadership, social and ethical values, and educational policy analysis.

Basic Education Reform in China: Untangling the Story of Success

Chengzhi Wang
Princeton University

Quanhua Zhou
University of Arizona

Abstract
China's recent basic education reform followed and, in a certain way, imitated its economic reform. The economic reform merged the experimental dual (planned and market) price systems into a free market economy and yielded phenomenal success. Basic education reform, however, has not succeeded in transforming the introductory dual-track (key school and regular school) systems into a universal one. This article briefly examines the general process and outcomes of basic education reform. It discusses the following questions: Is basic education reform also a story of success? What significant lessons can the Chinese reform experience offer to other comparable developing countries?

Introduction

The reform of basic education (which includes primary and junior secondary schooling) in China from the middle 1980s has not completely severed it from Maoist popular education. The post-Mao reform policy makers have never discarded the tradition of localization and community participation. In contrast to Maoist egalitarian schooling, however, school or pupil tracking (typically represented by key vs. regular schools) has been promoted in pursuit of economic efficiency in post-Mao educational changes and reforms.

This article presents a brief examination of the general process and outcomes of basic education reform. We first summarize economic reform and basic education reform, in particular their significant similarities and differences in terms of process and results. We then explain the success of basic education reform using three perspectives, namely, 1) the three matters/solutions, 2) contingency theory, and 3) the 3-C framework. Next, we analyze the price that China has paid for the success of education reform. Finally, we conclude that what the Chinese experience can offer to other developing countries is just what other countries have offered to China: erosion of traditions and westernization of schooling.

Economic Reform

Chinese economic reform is a unique process. From a price perspective, in the early 1980s, the government acquiesced to the coexistence of central planned production and market pricing. In 1985, transactions based on market prices outside the state plan won legal sanction. Gradual decontrol of consumer goods prices steadily brought most consumer goods into a market price system (Naughton, 1995; Riskin, 1987). In 1991, the Central Committee of the Communist Party called for elimination of the dual-track system and boldly recommended a gradual shift to a market system. One year later, the National People's Congress declared that the objective of reform was a "socialist market economy with all stress on the free market" (Naughton, 1995, p. 288). The government then unambiguously embraced the free market economy and began systematically dismantling the outdated command plan economic structure.

However, the economic reform was not strategically planned. In other words, it was initiated without a strategy. Yet, "a limited number of crucial government decisions and commitments were required in order to allow reform to develop. In certain periods, policymakers acted as if they had a commitment to a specific reform strategy" (Naughton, 1995, p. 7). In the process of the reform as a whole, "what is most striking is the succession of incremental, steadily accumulating measures of economic reform that have gradually transformed the economy in a fundamental way" (Naughton, 1995, p. 20).

No doubt, the two decades of economic reform resulted in increasing income inequality as documented in the rich research literature studying the reform. Yet, the growth of an income gap is not peculiar to China. It is a worldwide phenomenon observed in both developed countries such as the U.S. and all transitional countries in recent decades. Furthermore, in the case of China, the extent of income inequality and its underlying causes are still very far from clear (Bramall, 2001). In terms of most important indicators, the Chinese economic reform has been a success (Maddison, 1998; Naughton, 1995).

Basic Education Reform

The economic reform brought about reform in the education sector, particularly basic education. Just as economic reform initially allowed planned and market price systems, basic education changes embraced dual track schooling systems in the late 1970s and early 1980s. Key schools (to be explained later) were introduced and developed in each province, prefecture, county, township, or even village to admit the better-performing students and highly qualified teachers in each respective jurisdiction. In the basic education reform in the middle 1980s, both key and regular schooling systems were expanded greatly (Lewin et al., 1994), but emphasis was placed on the former. Financial, physical, and human resources that were supposed to otherwise be distributed equitably for all schools and students were concentrated on key schools. In a regular school that does not boast key status, key/fast classes were developed for better-performing students through utilizing the school's concentrated resources. In essence, the dual-track schooling system is a bifurcated educational system with a small sector of key schools for the elite and a large sector of regular schools for the masses (Rosen, 1985, 1987). This educational differentiation is ubiquitous in the Chinese educational hierarchy from kindergartens to universities.

Key schools were developed to achieve two goals. The first was to produce maximum educational returns in the shortest time, particularly to produce more qualified graduates for higher-level institutions in order to meet the immediate growing manpower demands. The second was to serve as a teaching and learning model for regular schools, so as to improve the overall quality of the whole basic education sector (Rosen, 1985, 1987). If the former goal has been partly realized by selectively promoting students to higher level institutions, particularly a very small percentage of students to colleges and universities, the second too ambitious goal has never been realized. In consequence, the diploma disease that hit other developing countries and spread in Chinese education reform in the 1960s (Unger, 1980) has resurfaced during the reform period under study here (Unger, 1980; Pepper, 1987, 1990, 1996, 1997).

In the 1990s, the college enrollment rate was steadily rising (Wang, 2000). Since the second half of 1990s, the exam-oriented schooling associated with the dual track system was considered too counterproductive to reform objectives, particularly universalization of 9-year schooling. The central educational authorities decided to de-emphasize the elite track. Education policy makers explicitly required that all primary and secondary schools should admit students in their neighborhoods and communities. In 1996, Li Lanqing, Vice Premier in charge of education, declared: "We must, from now on, no longer promote key middle schools or continue contributing all of our human, physical, and financial resources and all of our subsidies and donations into such schools" (Li, 1997). Merging the educational dual tracks, however, turned out to be a far more difficult and lengthy enterprise for a variety of political, socioeconomic, and educational reasons. The reintegration of the two tracks has yet to be completed.

Nevertheless, to a great extent, the assessment of economic reform discussed earlier also applies to education reform, particularly basic education reform. There was no clear strategy but instead feverish negation of Maoist popular education at the outset of schooling reform. The compulsory education law enacted in 1986 served more to declare the importance of universal education than to enforce the 9-year compulsory schooling at the initial stage. The objective of educational finance reform (notably the national promotion and standardization of education surcharges in the early 1990s following the termination of the state's provision of financial resources to basic education) was for the central government to fend for itself in the larger finance reform (Wong, 1997). The educational practices and realities for both women and national minorities have been oftentimes contradictory with the declared policies and priorities

(Rosen, 1992, 1995).

Notwithstanding, the basic education system has experienced significant transformation in the two decades of reform. First and foremost, the reform goals of 9-year compulsory schooling and literacy have been largely realized. Second, resource mobilization has resulted in relatively adequate financial resources for 9-year compulsory schooling. Lastly, the educational landscape of diversity beyond the Ministry of Education (MOE) system has taken shape, particularly along with the significant expansion of private schools and NGO-sponsored Hope Schools and the indispensable contribution of Maoist *minban* ("people-managed," or community-supported) teachers in rural areas. Admittedly, governmental decisions were crucial in transforming the basic education system in a fundamental way. These decisions included moves to decentralize educational governance, universalize 9-year schooling and improve literacy, diversify educational financing, and enforce education taxation by garnering resources from communities and households. In the final analysis, China's basic education reform has been, in short, a success.

Explaining Success

Indeed, significant progress in national education attainment and schooling effectiveness has been achieved in China in terms of what Levin and Lockheed (1993) called three matters. The relative efficiency of education can result from, first, growing participation (enrollment, completion, and achievement), second, more effectiveness (less dropout and repetition, and positive learning result), and third, increasing resources (more expenditures per student, annual recurrent public educational expenditures, qualified teachers, facilities, textbooks, and others) (Levin and Lockheed, 1993, pp. 1-19). Generally, the three matters have been, to a great extent, well addressed in China in the reform era, especially in the late 1990s (MOE Department of Planning and Development, 1998). In a certain sense, Levin and Lockheed's (1993) prescription of "three solutions" for creating effective schooling in developing countries, namely, "basic inputs," "facilitating conditions," and the "will to act" of government and communities (p. 13), have been made all available in the Chinese case of basic education (Ahmed et al., 1991; Lewin, et al., 1994; Tsang, 1996).

In the past two decades, high-profile publicity campaigns for reform and expansion of basic education repeatedly swept the whole nation. The annual evaluation of governmental officials at each administrative level included schooling effectiveness as an important performance indicator. The central educational authorities issued general standards and requirements for primary and secondary education in terms of school building, classrooms, textbooks, teaching aids, playgrounds, toilets, drinking water, and others. Due to regional differences in terms of resources, there are gaps between urban and rural schools, between schools in developed areas and underdeveloped areas, and between schools in the *Han* majority areas and schools in minority areas. However, in the poorest school communities, the Ministry of Education minimum requirements were "*yiwu liangyou*" (one "have-not" and two "haves"). Namely, no school should have dangerous school buildings or facilities; every school must have classroom buildings and every student must have a desk and stool in the classroom (Cheng, 1993, pp. 46-62; West, 1997, pp. 214-246). Because culturally students are entitled to textbooks and, administratively, textbooks have always been centralized and guaranteed even for the poorest schools, the "have" of textbooks is not specified in the minimum requirements (Cheng, 1993, pp. 46-62). The "*ba peitao*" (eight supplements) have also been advocated in the Ministry of Education guidelines. Namely, every primary or secondary school must be supplemented or outfitted with a school gate for safety, a garden, necessary toilets, a cultural (recreation) room, a laboratory, a library, and a sports ground. Except for the schools in poorest rural areas, which may not have a laboratory or library, most schools have met the most important requirements. In addition, "*santong*" (three connections) have to be realized. Namely, water, electricity, and road linkups to every school must be guaranteed. Except for a few schools in locations with extremely adverse conditions that preclude electricity connections most schools have realized the three connections (West, 1997, pp. 214-246).

Second, the story of success can also be explained by using contingency theory, which partly holds that there is no single best decision-making approach (Tarter & Hoy, 1998). The difficulties of implementing education reform in developing countries range from complexity of reform proposals, unpredictability of education reforms, inappropriate management strategies, a failure to focus on school level changes, and a failure to assess the capacity of organizations to manage innovation (Rondinelli et al., 1990, pp. 11-15). These difficulties need to be addressed by using the contingency approach in the planning and implementation of education reforms. The application of contingency theory overcomes the weaknesses of conventional planning and recognizes the central position of people. It takes unstable and uncertain external environments into consideration by targeting both routine and innovative tasks. Equally important, it tolerates ambiguity and risk (Rondinelli et al., 1990).

The policy decision making process of Chinese education reform is strongly impacted by external environments and factors. The controversial dual-track scheme that involves every school and every student, intends to address both routine and innovative tasks of education. More often than not the education reform legislation and policies are ambiguous. The policy process, particularly the implementation of governance decentralization and financing diversification, has been exploitative of households and communities and, sometimes, misguided. However, despite groping for stones to cross the river, both policy makers and implementers are fully aware of the minimum goals that they have to reach: to universalize compulsory basic education and to increase literacy. Peasants, rural communities, grassroots organizations, and local governments have played a pivotal role in adopting policies, implementing programs, and providing resources to reach the fundamental targets of compulsory schooling.

"Indeed, the Chinese reform kept evolving in ways that policymakers didn't anticipate, and they had to scramble to

205

catch up with the changes they had unleashed" (Naughton, 1995, p. 23). In the process of reform, each step taken depended on the previous step. From the perspective of contingency theory, "since by significant resilience as well, such an approach might be admired as the strategy of not having a strategy, or as we might say, of 'muddling through'" (Naughton, 1995, p. 7).

We choose to term the continuous and seemingly endless but evolving reform process following the first initiated educational change and reform as "education re-reform." Education re-reform is a unique phenomenon in the Chinese education sector. In Chinese economic development, "reform proceeds by a series of feedback loops—reform begets further reform" (Naughton, 1995, p. 320). In education, decentralization and de-politicization featured with the reform in the 1980s were ignited, simply put, to redress hyper-politicization in the Maoist era. The reform begat changes with growing tendency of centralization or re-centralization in the 1990s. The reform of financing diversification was first intended to redress over-centralization in the early 1970s. The diversification and resource mobilization also begat reform of more central and upper level interventions in the late 1990s. China's education reform could not proceed fruitfully without a number of timely state decisions and interventions. The continuous education re-reform may result in, what Lampton (1987) calls, incongruencies between policy intentions and outcomes (p. 8). In fact, it has done so in many aspects. Nonetheless, "the dynamics of the process create opportunities for pro-reform leaders to push the reform forward" (Naughton, 1995, p. 320). Also from the perspective of contingency theory, the dynamics of seemingly irregular and repeated re-reforms are reshaping the Chinese education system to a new stage of development, with numerous mistakes and failures of course, but with more exciting stories of success often unheard in other comparable developing countries.

Lastly, the simplistic, but interesting 3-C (consistency, connectedness, and culture fit) frame, which some used to evaluate education reform in Hong Kong (Dowson et al., 2000), may also be used to explain the complicated case of educational development in China's mainland. According to Dowson et al. (2000), consistency refers to how the thrust of the reforms and reform components are interpreted. That is: are the reforms consistent, or do they confuse educators through proposing apparently contradictory purposes? Connectedness refers to whether reform components are linked in terms of what they are trying to achieve and how they are achieved. Is the huge array of quality reforms coherently connected to each other? Cultural fit refers to whether the reforms and reform components are appropriate given the traditional culture and the context of educational institutions. Are the thrusts of the reforms culturally appropriate?

The examination of the basic education reform in China's mainland has provided negative answers with significant evidence, particularly to questions about the first two indicators. For example, educational differentiation and segregation are very difficult to overcome and the misguided dual track system is hard to merge into an appropriate equitable educational system under the efficiency-oriented environments of the free market economy. Policy vagueness, inarticulation, and confusion are notoriously ubiquitous in the Chinese education system, in particular in the four domains of the policy process of formulation and consultation, adoption, implementation, and monitoring (NEPI, 1992, 1993) at the national and provincial levels. There are severe mismatches and contradictions between compulsory education policies and practices. Teachers in economically underdeveloped areas are offered low and late pay, but are urged by the state to play a greater role in improving school effectiveness and universalizing compulsory education.

However, basic education reforms also display major threads of consistency, connectedness, and cultural fit. In particular, the mobilization of resources, which is very important for successful expansion of basic education but more difficult to practice in many other developing countries, is fitting in with the traditional culture that emphasizes support to schooling. Ahmed et al. (1991) observed: "Policy formulation is clearly a central function, but the devolution of financial responsibility for basic education to the local level has led to considerable de facto flexibility in the application, adaptation and interpretation of policies" (p. 162). This flexibility is reflected in, for example, locally initiated extensive education surcharges and levies. The extensive, sometimes rampant, fees and surcharges have incremented significant financial burdens for communities and peasants in economically underdeveloped school communities. However, the practice of resource mobilization (Tsang, 1994, 1996) also results in a higher level of creativity, pragmatism, and productivity, ensuring the generally adequate financial resources available for almost every school. Compared with basic education development in other developing countries such as India (PROBE Team, 1999), one may argue that the Chinese basic education reforms evidence a greater degree of continuity and consistency of policies in the overall positively steered process and environments (Ahmed et al., 1991, pp. 162-174). In other words, the lack of consistency, connectedness, and cultural "fit" of basic education reforms does exist; but, it is not as significant as in other comparable developing countries.

The Price of Success: Erosion of Tradition

The relative success of education reform has not been achieved without a price: the erosion of traditions, as well as the erosion of equities. Traditions are transforming into a new uncertain combination in the process of reform. Pepper (1987) insightfully examined Chinese education in the 1980s:

> "Three diverse traditions came together in Chinese education during the 1950s, in an uncertain combination that has yet to be fully reconciled thirty years later. The tradition that the Chinese Communist Party inherited from the Republican era was itself an amalgam of modern Western-inspired learning grafted on an ancient Confucian base. The second tradition the Chinese

206

Communists brought with them from their own recent experience as leaders of the rural Border Region governments in the 1930s and 1940s. The third tradition was introduced into China in the 1950s, when the new Communist government embarked on an ambitious attempt to learn from the Soviet Union. The influence of each of these three traditions can still be seen in Chinese education, their outlines now firmly etched in the public mind and in official discourse by the volatile combination they have produced." (p. 185)

After two decades of reform, the first tradition is diminishing along with the passing away of the first generation Western-inspired political and intellectual leaders. The ancient Confucian base is shrinking even further with the powerful intrusion of modernization and forceful encroachment of free market economy. The second tradition, the indigenous Communist tradition, has suffered most. The drive for modernization inherently undermined the traditional society; and the transformations, natural and man-made, shattered the rhythms of the past (MacFarquhar, 1987, p. 542). With increasing merciless globalization, the last traces of this indigenous tradition may be expected to be disposed of completely in the foreseeable future. The third tradition demised with the Soviet Union. This external tradition transformed and recreated itself into the American tradition. In other words, the United States replaced the Soviet Union to establish the Americanized tradition and exercise its growing impacts. Although the lessons learned from the Soviet story have made China cautious in relearning, the United States has become the exemplary model of China in the reform era.

The new combination of the transformed traditions has now resulted in no traditions, or more precisely, the imported traditions of the West and of America. Like students in other developing countries, school children in China are also "alienated from the past" (Rodriguez, p. 174). The cost of "losing the past" and losing identity is not negligible. People feel their education has put them in a societal black hole (Kazmi, 1989, pp. 171-177).

Conclusion

Notwithstanding the relevance of the reforms of governance and financing of Chinese basic education to other developing countries, the concepts of "success" and "failure" of policies and practices are themselves actually illusory (Lampton, 1987, p. 5). So are the concepts of educational "success" and "failure" for the Chinese case of basic education reform.

From the erosion and transformation of traditions, as well as from the growing inequalities and inequities in most developing countries, what the Chinese case can offer to other developing countries is exactly what other developing countries can teach China: for better or worse, the traditions are eroded in education reforms. The state becomes the fragile state (Fuller, 1990). Chinese schools, as well as other Third World schools, are built in the Western way.

Acknowledgements

References

Ahmed, M. with Cheng, K. M., Jalaluddin, A. K., Ramachandran, K. (1991). Basic education and national development. New York: UNICEF.

Bramall, C. (2001). The quality of China's household income surveys. *China Quarterly, 167*, September, 689-705.

Cheng, K. M. (1993). Zhongguo dalu jiaoyu shikuang (The true situation of education in mainland China). Taipei, Taiwan: Taiwan Commercial Press.

Dowson, C., Bodycott, P., Walker, A., & Coniam, D. (2000). Education reform in Hong Kong: Issues of consistency, connectedness and culture. *Education Policy Analysis Archives, 8*(24). Retrieved November 18, 2001 from http://epaa.asu.edu/epaa/v8n24.html.

Fuller, B. (1990). *Growing up modern: The western state builds third world schools.* New York: Routledge.

Kazmi, Y. (1989), *An ontological notion of tradition and education.* Ph.D. Dissertation. Urbana, IL: University of Illinois at Urbana-Champaign.

Lampton, D. M. (1987). *Policy implementation in post-Mao China.* Berkeley, CA: University of California Press.

Levin, H. M. & Lockheed, M. E. (1993). *Effective schools in developing countries.* Washington, D.C.: Falmer Press.

Lewin, K., Little A., Xu, H., & Zheng, J. (1994). *Educational innovation in China: Tracing the impact of the 1985 reform*. Harlow, UK: Longman.

Li, L. (1997). The fundamental mission of basic education is to enhance the quality of the entire nation (Speech by Vice Premier Li Lanqing during an inspection of quality education at Miluo). *China Education and Society, 30*(6), November-December, 29-36.

MacFarquhar, R. (1987). Epilogue. In D. Twitchett & J. K. Fairbank (Eds.), *The Cambridge history of China* (Volume 14). New York: Cambridge University Press

Maddison, A. (1998). Chinese economic performance in the long run. Paris: Development Centre of OECD

MOE Department of Planning and Development (1998). Educational statistics yearbook of China. Beijing: People's Education Press.

Naughton, B. (1995). *Growing out of the plan*. New York: Cambridge University Press.

NEPI (1992). *Governance and administration*. Cape Town, South Africa: Oxford University Press

NEPI (1993). Toward democratic governance of education in a new South Africa, *Forum for Advancing Basic Education and Literacy*, May, 10-11.

Pepper, S. (1987). Education for the new order. In D. Twitchett & J. K. Fairbank (Eds.), *The Cambridge history of China: The People's Republic* (Volume 14). New York: Cambridge University Press

Pepper, S. (1990). China's education reform in the 1980s: Policies, issues and historical perspective. Berkeley, CA: Institute of East Asian Studies, University of California.

Pepper, S. (1996). *Radicalism and education reform in twentieth-century China*. New York: Cambridge University Press.

Pepper, S. (1997). Guest editor's introduction. *China Education and Society, 30*(2), March-April, 7-10.

PROBE Team (1999). Public report on basic education in India. New Delhi, India: Oxford University Press.

Riskin, C. (1987). *China's political economy*. New York: Oxford University Press

Rodriguez, R. (1983). Hunger of memory. New York: Bantam Books. In Kazmi, Y. (1989), *An ontological notion of tradition and education*. Ph.D. Dissertation. Urbana, IL: University of Illinois at Urbana-Champaign.

Rondinelli, D. A., Middleton J., & Verspoor A. M. (1990). *Planning education reforms in developing countries*. Durham, NC: Duke University Press.

Rosen, S. (1985). Recentralization, decentralization, and rationalization. *Modern China, 11*(3), 301-346.

Rosen, S. (1987). Restoring key secondary schools in post-Mao China: The politics of competition and educational quality. In D. M. Lampton (Ed.), *Policy implementation in post-Mao China* (pp. 321-353). Berkeley, CA: University of California Press.

Rosen, S. (1992). Women, education and modernization. In R. Hayhoe (Ed.), *Education and modernization in Chinese experience* (pp. 255-284). New York: Pergamon.

Rosen, S. (1995). Women and reform in China. In G. A. Postiglione & W. O. Lee (Eds.) *Social change and educational development* (pp. 120-129). Hong Kong: Center of Asian Studies, the University of Hong Kong.

Tarter, J. C. & Hoy, W. K. (1998). Toward a contingency theory of decision making. *Journal of Educational Administration, 36*(3-4), 212-228.

Tsang, M. C. (1994). Costs of education in China: Issues of resources mobilization, equality, equity and efficiency. *Education Economics, 2*(3), 287-312.

Tsang, M. C. (1996). Financial reform of basic education in China. *Economics of Education Review, 15*(4), 423-444.

Unger, J. (1980). Bending the school ladder: The failure of Chinese educational reform in the 1960s. *Comparative Education Review*, June, 221-237.

Wang, C. (2000). From manpower supply to economic revival: Governance and financing of Chinese higher education, *Education Policy Analysis Archives, 8*(26). Retrieved November 15, 2001 from http://epaa.asu.edu/epaa/v8n26.html.

Wong, C. P. W. (1997). *Financing local government in the People's Republic of China*. New York: Oxford University Press.

About the Authors

Chengzhi Wang
061 Wallace Hall
Princeton University
Princeton, NJ 08544

E-mail: chengzhi@princeton.edu

Chengzhi Wang did his doctorate work in educational policy studies at the University of Illinois at Urbana-Champaign. He taught English education at a normal university in China and multicultural education at a teaching university in Illinois. He now works as an academic professional at the Princeton University.

Quanhua Zhou
College of Education, Room 321
University of Arizona
Tucson, AZ 85721-0069

E-mail: qzhou@u.arizona.edu

Quanhua Zhou is a doctoral student at the Center for the Study of Higher Education at the University of Arizona. He earned a Master's degree in International Relations and worked in the administration of Peking University for several years. His research interests include higher education finance and educational policy.

EPAA Editorial Board

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Quantifying Quality: What Can the *U.S. News and World Report* Rankings Tell us About the Quality of Higher Education?

Marguerite Clarke
Boston College

Abstract

Since their first appearance in 1983, the *U.S. News and World Report* rankings of colleges and graduate schools have generated much discussion and debate, from some declaring them among the best rankings ever published to others describing them as shallow, inaccurate, and even dangerous. The research presented here addresses two of the most common criticisms of the methodology used to produce these rankings. In particular, this study answers the following questions: What is the extent of change in *U.S. News'* ranking formulas across years and what are the implications for interpreting shifts in a school's rank over time? How precise is the overall score that *U.S. News* uses to rank schools and what are the implications for assigning schools to discrete ranks? Findings confirm critic's concerns in each of these areas, particularly in relation to the ranking of graduate schools of education. Based on these results, five recommendations are made for improving the interpretability and usefulness of the rankings.

Introduction

Every year, *U.S. News and World Report's* (*U.S. News*) rankings of the academic quality of colleges and graduate schools hit the newsstands (Note 1). Their arrival brings delight to some and dismay to others, depending on whether their institution rose or fell in the quality ratings. An improved ranking can lead to increased donations from proud alumni and more and better qualified students in next year's applicant pool (Monks and Enhrenberg, 1999). A fall can lead to tighter alignment of institutional benchmarks and goals with ranking criteria and pressure on admissions staff to bring in "better" applicants (Mufson, 1999). All the while, a question goes unanswered: What do these rankings really tell us about the quality of higher education?

As a step toward answering this question, I examine two common criticisms of the methodology that *U.S. News* uses to rank colleges and graduate schools. These are: (1) constant changes to the formula make it impossible to interpret yearly shifts in a school's rank in terms of change in its relative academic quality (Levin, 1999; Pellegrini, 1999), and (2) the score used to assign schools to ranks is overly precise, creating a vertical column where a group might more properly exist (Machung, 1998; Smetanka, 1998). The first section of this article gives a brief introduction to the *U.S. News* rankings as well as the questions addressed by this study. The next section outlines the methodology used to answer these questions and the results of the analyses. The final section presents conclusions and recommendations.

Before proceeding, a caveat is in order. While many have questioned the overall concept of academic quality rankings as well as the validity of the different indicators and weights used, I suspend judgment on these issues to focus on the extent to which methodological problems may impact the interpretation of the *U.S. News* rankings.

Background on the *U.S. News* Rankings

U.S. News published its first rankings of the academic quality of colleges in 1983, the same year that the National Commission on Excellence in Education released *A Nation at Risk,* its influential report blasting the quality of education in America. Based on a survey of college presidents, the magazine listed Stanford, Harvard, and Yale as the top three national universities and Amherst, Swarthmore, and Williams as the top three national liberal arts colleges. By 1987, *U.S. News* had moved to a multidimensional approach, weighting and combining information on faculty accomplishments, student achievements, and institutional academic resources to produce an overall score on which to rank colleges. Rankings of graduate schools of business, engineering, law, and medicine/primary-care also appeared in this year and used a similar weight-and-sum approach (rankings of graduate schools of education did not appear until 1994).

The most recent rankings still use this basic approach. At the undergraduate level, schools are categorized by mission and region (e.g., national universities, national liberal arts colleges, regional universities, and regional liberal arts colleges). Up to sixteen pieces of information are collected on schools in each category, including academic reputation; freshmen retention and graduation rates; average test scores for entering students; per-student spending; and alumni-giving rate. These indicators are standardized, weighted, and summed to produce an overall score on which to rank schools in each category against their peers.

At the graduate level, schools are categorized by type—business, education, engineering, law, and medicine/primary-care. Depending on the type of school, data on up to fourteen indicators—including test scores, research expenditures, graduate employment rates, and reputation—are collected. Similar to the undergraduate rankings, the indicators are standardized, weighted, and summed to produce an overall score on which to rank schools in each category against their peers. Detailed information on the indicators and methodology that *U.S. News* uses to rank colleges and graduate schools is found in Appendix A. (Note 2)

Criticisms of the *U.S. News* Rankings

Almost two decades after their first publication, the college and graduate school rankings are among *U.S. News'* top issues in terms of sales generated (K. Crocker, personal communication, March 19, 1999). This demand has made them the focus of much criticism and debate, especially among the institutions that are the subject of the rankings. In addition to questioning the overall concept of ranking higher education institutions, much criticism has focused on the methodology used to produce the rankings. Gerhard Casper, then President of Stanford University, focused on some of these methodological concerns in a letter of protest he wrote to the editor of *U.S. News* in 1996:

> Could there not, though, at least be a move toward greater honesty with, and service to, your readers by moving away from the false precision? Could you not do away with rank ordering and overall scores, thus admitting that the method is not nearly that precise and that the difference between #1 and #2 - indeed, between #1 and #10 - may be statistically insignificant? Could you not, instead of tinkering to "perfect" the weightings and formulas, question the basic premise? Could you not admit that quality may not be truly quantifiable, and that some of the data you use are not even truly available (e.g., many high schools do not report whether their graduates are in the top 10% of their class)? Parents are confused and looking for guidance on the best choice for their particular child and the best investment of their hard-earned money. Your demonstrated record gives me hope that you can begin to lead the way away from football-ranking mentality and toward helping to inform, rather than mislead, your readers. (Note 3)

Casper's questions about the "football ranking mentality" employed by *U.S. News* go to the heart of the debate over college and graduate school rankings. If, as Casper states, "the difference between #1 and #2 - indeed, between #1 and #10 - may be statistically insignificant," what are the implications for the way in which the overall scores for schools are used to put them in rank order? In addition, if the weights and formula are constantly being "tinkered" with, how should one then interpret change in a school's rank from year to year?

Others have voiced these methodological concerns. In particular, critics have noted that yearly formula changes make it almost impossible to interpret shifts in a school's rank in terms of change in its relative academic quality: a college that is ranked 4th one year and 7th the next may have had no change in its performance relative to other schools, yet still have moved because of changes in the ranking methodology (Levin, 1999; Machung, 1998; Pellegrini, 1999). *U.S. News'* response to this issue has been that they prefer to make incremental changes every year to produce the "best possible rankings" than to use the same indicators every year to facilitate precise year-to-year comparisons.

Critics have also pointed out that the use of overall scores to rank schools magnifies small—and often insignificant—differences among schools, and that small changes by the school or the magazine can move a college half a dozen places up or down the ranking list (Crenshaw, 1999). *U.S. News* acknowledged this issue in 1998 when it began

rounding overall scores to the nearest whole number in recognition, the editors noted, of the fact that small differences after the decimal point may reflect non-significant differences between schools (Thompson and Morse, 1998). Subsequently, the number of schools tied for overall score (and thus rank) increased dramatically.

While much criticism and debate has focused on the methodology used to produce the rankings, the majority of research has focused on the extent to which the rankings are used by students and parents (e.g., Art and Science Group, 1995; McDonough, Antonio, Walpole, and Perez, 1998) or their effect on institutions (e.g., Monks and Ehrenberg, 1999). The research presented here addresses the two methodological concerns outlined above. In particular, this study answers the following questions:

1. What is the extent of change in U.S. News' ranking formulas across years and what are the implications for interpreting shifts in a school's rank over time?
2. How precise is the overall score that U.S. News uses to rank schools and what are the implications for assigning schools to discrete ranks?

Methods and Results

Tracking Changes in Ranking Formulas across Years

In order to gauge the extent of change in the U.S. News ranking formulas over time, year-to-year changes to the indicators used in each formula were tracked across rankings published between 1995 and 2000 inclusive. Four types of changes were identified and tracked over this six-year period: changes in the weight assigned to an indicator; the removal of an indicator from a formula; the addition of an indicator to a formula; and, changes in an indicator's definition or methodology. Rankings examined included business, education, engineering, law, and medicine/primary-care at the graduate level and national university and national liberal arts college at the undergraduate.

Changes in weights, methodology, and the addition or removal of indicators were generally easy to track, although it was not possible to fully track changes in weights at the undergraduate level as this information was not included until the 1998 edition of the guidebook. Changes in indicator definition were harder to identify as the wording for a definition could differ from one year to the next, while the underlying meaning might not. The following rule was used to identify an indicator definition change:

1. The new wording must contain additional detail such as a date, money amount, percent, or other precise information not previously stated or implied.
2. If the new wording does not include such detail, it should be recognized as changed by U.S. News in the guidebook text.

Analyses focused on the types of changes that were made to the formula for each ranking, the total number of these changes across time, the proportion of non-change in each ranking formula, and the extent to which the amount of change in a ranking formula was related to the amount of movement in the relative ranks for schools in that ranking across the same time period.

Table 1 summarizes changes in the indicators used for each ranking from 1995 to 2000. The number of changes for each ranking, by type and overall, is shown in columns two through eight. The national university and national liberal arts college changes are shown in one column as they use the same formula. The final column in Table 1 reflects the total number of changes across all seven rankings (i.e., business, education, engineering, law, medical, national university/liberal arts, and primary care), again broken down by type.

Table 1
Changes in U.S. News Ranking Indicators, 1995-2000

	Business	Education	Engineering	Law	Medical	National University/ Liberal Arts	Primary Care	Total
Definition/ Methodology	4 (50)*	4 (67)	3 (37.5)	10 (72)	4 (100)	4 (50)	3 (60)	32 (60)
Weight	3 (37.5)	2 (33)	3 (37.5)	1 (7)	0	2 (25)	2 (40)	13 (25)
Addition	0 (0)	0 (0)	1 (12.5)	1 (7)	0	1 (12.5)	0	3 (6)
Removal	1 (12.5)	0 (0)	1 (12.5)	2 (14)	0	1 (12.5)	0	5 (9)
Total	8 (100)	6 (100)	8 (100)	14 (100)	4 (100)	8 (100)	5 (100)	53 (100)

213

*Column percentages are in parentheses.

Most changes were weight or definition/methodology changes, comprising 85 percent of all changes occurring over the six editions. Very few indicators were added to or removed from the ranking formulas, suggesting that *U.S. News* generally retained the same set of indicators for each ranking, but consistently refined and redefined these indicators over the years. (Of course, this redefining process can also change an indicator substantially).

The rate of change varied widely across rankings. While most rankings averaged between 6 and 8 formula changes over the six editions, the law rankings experienced 14 and the medical rankings only 4 changes over the same period. Several reasons account for the larger number of changes in the law ranking's indicators, including *U.S. News'* responses to the complaints of law schools (who tend to complain more than other schools) and the release of new types of quality-related information by the American Bar Association.

While a ranking (e.g., the law rankings) may have experienced a large number of changes relative to other rankings, these changes may be concentrated in a small group of indicators that are constantly being refined. Different rankings of schools also use different numbers of indicators to compute their overall score, and thus two rankings that experience the same types and number of changes may differ in the number of indicators left unchanged overall. Figure 1 shows the proportion of unchanged indicators for each ranking between 1995 and 2000 inclusive.
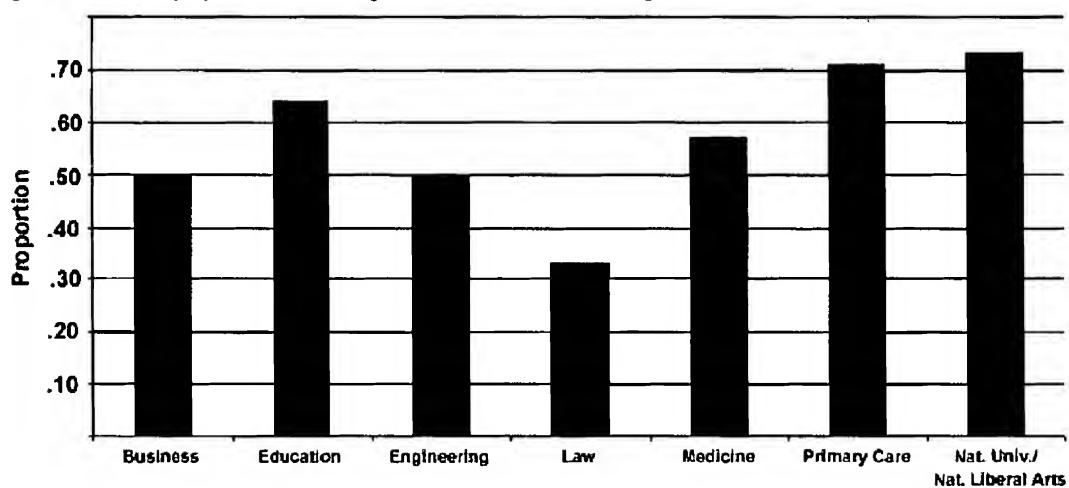
Figure 1. Proportion of Indicators Remaining Unchanged in Each *US News* Ranking, 1995-2000.

The undergraduate rankings (both national university and national liberal arts college) have the largest proportion (.73 approximately) of unchanged indicators. In contrast, only about one third of the law school indicators remained unchanged. For most rankings, about half to two thirds of the indicators remained unchanged over the six editions. This suggests that while it may not be always possible to interpret changes in a school's overall rank across years, it is possible to track performance on individual indicators that have remained unchanged across the years. Most of the unchanged indicators are related to selectivity (e.g., test scores and the proportion of applicants accepted into the program) and institutional resources (e.g., student-faculty ratios).

In Table 2, an X indicates when it is possible to make cross-year comparisons for a ranking. The criteria used to make this determination include the four types of indicator changes discussed above as well as more general formula changes. The latter occurred twice over the six editions examined here: In 1998 when overall scores were rounded to the nearest whole number, and in 1999 when a school's performance on each indicator was standardized before obtaining the overall rank score. While it was not possible to make cross-year comparisons for most rankings over the six years, the last column in Table 2 suggests that the ranking formulas may be stabilizing. Between 1999 and 2000, there were no changes in the formulas used to rank schools of education, engineering, law, and medicine, suggesting that change in a school's rank between 1999 and 2000 could be interpreted in terms of change in its relative academic quality.

Table 2
Ability to Make Comparisons Across Years for a Ranking, 1995-2000

Ranking	1995-1996	1996-1997	1997-1998	1998-1999	1999-2000
Business	X				
Education					X
Engineering					X
Law					X

Medical	X	X			X
National Liberal Arts					
National University					
Primary Care	X				X

It is important to remember that even when a formula appears to remain stable across years, there can still be difficulties with cross-year interpretation of ranks. This is due to problems with the accuracy of the information obtained and critics have pointed out several errors that have arisen due to mistakes (both accidental and deliberate) in reporting by institutions, and due to the differing ways in which schools compute figures for certain indicators (Machung, 1998, Smetanka, 1998, Stecklow, 1995, Wright, 1990-91). *U.S. News* has tried to reduce the error introduced by these practices by cross-checking data sent in by schools with data collected by debt-rating agencies, investors and national organizations such as the National Collegiate Athletic Association, and tightening up their survey questions, but issues still remain.

The final stage of the comparability analysis examined the extent to which the amount of change in a ranking formula is related to the amount of movement in schools' ranks for that ranking across years. Table 3 shows the correlation (r) between the 1995 and 2000 ranks for the top-fifty schools in each ranking in 1995.

Table 3
Correlation between 1995 and 2000 Ranks for the
Top-Fifty Schools in 1995, By Ranking

Ranking	Correlation (r)
Business	.89
Education	.72
Engineering	.88
Law	.92
Medicine	.88
National Universities	.95
National Liberal Arts College	.94
Primary Care	.08

There is no definite relationship between the amount of change in the indicators for a ranking and the correlation between the 1995 and 2000 ranks for the top-fifty ranked schools in 1995. For example, while law schools experienced the most change in their indicators over the six editions of *U.S. News*, there was not much difference (r = .92) in the rank ordering of the top-fifty law schools in 1995 and their ordering in 2000. While varying amounts of change was experienced in the indicators used for the other rankings, they still show a high degree of similarity (with r's between .88 and .95) in the rank ordering of their top 50 schools in 1995 and 2000. The main exceptions to this are the education (r = .72) and primary-care (r = .08) rankings. The low correlation between the primary-care rankings in 1995 and 2000 can be explained by changes in the population of schools that *U.S. News* included in these rankings during this time period. In contrast, the low (relative to the other rankings) correlation between the 1995 and 2000 ranks of the top-fifty schools of education in 1995 is linked to the fact that 16 of the top 50 schools in 1995 had experienced large changes in rank--of ten or more--by the 2000 edition. Table 4 shows the 16 schools of education. The first six schools all experienced a decline in rank, ranging from a drop of 10 places for the University of Southern California and the University of Iowa to a drop of 22 places for Syracuse University. The remaining schools all improved their rank since 1995. Improvement ranged from an increase of 10 places for the Rutgers University to a jump of 30 places for Arizona State University.

Table 4
Schools of Education with the Biggest Differences in
***U.S. News* Rank between 1995 and 2000[a]**

School	Rank						Change in Rank Between 1995 and 2000
	1995	1996	1997	1998	1999	2000	
University of Iowa	20	22	14	15	27	30	-10
University of Southern	23	27	26	30	31	33	-10

California							
University of Georgia	15	10	15	19	18	26	-11
SUNY-Buffalo	39	45	43	47	46	Not Ranked	At least -12
Boston University	31	37	32	43	Not Ranked	46	-15
Syracuse University	28	41	46	45	46	50	-22
Rutgers State University-New Brunswick	49	33	29	30	33	39	+10
University of Minnesota-Twin Cities	25	7	9	11	10	14	+11
University of Pittsburgh, Main Campus	44	Not Ranked	43	34	37	33	+11
Temple University	33	30	34	28	20	20	+13
George Washington University	45	39	37	30	34	30	+15
University of Michigan-Ann Arbor	22	9	8	6	8	7	+15
University of North Carolina-Chapel Hill	32	32	31	28	22	17	+15
University of Texas-Austin	27	19	12	13	11	12	+15
New York University	40	28	23	19	16	12	+28
Arizona State University-Main Campus	47	29	39	27	24	17	+30

[a]This table does not include schools that were not ranked in 1995 but appeared in the top 50 in the 2000 edition.

Cross-year data for the top-fifty schools in 1995 in other rankings were also examined to assess the extent to which similar movements in rank occurred (only data for the top 25 schools of medicine/primary-care and the top 40 national liberal arts colleges were available). Only nine business schools, one engineering school, eight law schools, no medical or primary-care schools, three national liberal arts colleges and two national universities differed by ten or more places in their 1995 and 2000 ranks.

It is not clear why there was more movement among schools of education compared to other types of schools. If changes in indicators (i.e., weight, definition, or other changes) are not responsible, movement could be due to changes in schools' performance on the indicators or errors or inconsistencies in the information reported by schools. Unfortunately, it is difficult to identify the real reasons for these movement patterns among schools of education over time, as well as why these differ from other rankings, as *U.S. News* did not print much information on schools' performance on the individual indicators until 1999.

Estimating Error or Uncertainty around the Overall Score

There is no universally agreed-upon set of information for creating academic quality rankings. Thus, various ranking efforts use indicators that differ in whole or in part from those used by others even when attempting to rank the same schools. It is not difficult to imagine that slight changes in the set of indicators used—such as the addition or removal of a single indicator—may move a school up or down a ranking, depending on how it performs on the indicator relative to other schools. To gauge the effect of slight changes in the set of indicators on the stability of the overall score and subsequent ranking for a school, a technique called jackknifing (Efron and Tibshirani, 1993) was applied to the data for the top-50 schools in each of the 2000 business, education, law, national liberal arts college, and national university rankings. (Note 4)

First, a baseline regression model was created for each of the rankings, with schools' overall scores as the dependent or outcome variable and the indicators used for each ranking as the independent or predictor variables. The overall fit of the model to the data was assessed in terms of the *adjusted* R Squared. Values of .9 and above were considered a good fit, meaning that the overall score predicted by the model for a school was highly correlated with the score produced by *U.S. News'* ranking formula, and that the regression model was an effective substitute for the weights-and-sum formula used by *U.S. News*. All models met this criterion, with *adjusted* Rs Squared varying between .99 for the national liberal arts college and national university models, .98 for the business school and law school models, and .95 for the education school model. (Note 5)

An approximation to a standard error for each school's overall score was obtained using the following formula (Efron and Tibshirani, 1993): (Note 6)

$$se_{jackknife} = \sqrt{\frac{n-1}{n}\left[\sum \hat{\theta_i} - \sum \frac{\hat{\theta_i}}{n}\right]^2}$$

where n is the number of regression models to be estimated and $\hat{\theta_i}$ is the predicted score for a school from the i regression model with one indicator removed.

The removal of one indicator at a time for the jackknife regression models did not seem to affect substantially the overall *adjusted* R Squared in most instances. For example, for each of the 9 models estimated using the law school data, the *adjusted R* Squared never varied by more than .01 from the *adjusted R* Squared for the overall model (i.e., .98), suggesting that the indicators are contributing fairly similar information to the estimation of the overall score. As a result, the jackknife standard errors are quite small, varying, in the case of law schools, from a low of .74 for the University of Michigan, Ann Arbor to a high of 3.06 for Harvard University. A similar range of standard error values was obtained for all rankings except for schools of education. The regression model for schools of education was not as robust to changes in indicators and the *adjusted R* Squared dropped considerably (by .13) when one indicator in particular—*Research Expenditure*—was removed. The resultant jackknife standard errors for schools of education are therefore quite large, varying from a low of 1.78 for Stanford University to a high of 11.98 for the University of Southern California.

Differences in the standard errors for individual schools are due to differences in how the removal of different indicators from the equation affects the prediction of their overall score. For schools that have large standard errors, the removal of certain indicators makes it much harder to predict the overall score they received from *U.S. News*. For school with smaller standard errors, the removal of indicators does not appreciably reduce the precision of estimation of their overall score. This suggests that schools are differentially affected by the presence or absence of certain indicators in terms of their overall score and subsequent rank.

This error estimate was then used in a t-test to assess the extent to which one school's overall score was significantly different from that of another. The t-test formula employed was: (Note 7)

$$t = \frac{x_1 - x_2}{\sqrt{(se_{x_1})^2 + (se_{x_2})^2}}$$

where x_1 is the overall score for school 1, x_2 is the overall score for school 2 and *se* denotes the standard error of the respective score.

The results of these comparisons are summarized in Tables 5 through 9 which are in the form of Excell spreadsheets. In each table, schools are ordered by their overall ranking score across the heading and down the rows. Read across the row for a school in order to compare its performance with the schools listed in the heading of the chart. The symbols indicate whether the overall score of the school in the row is significantly lower than that of the comparison school in the heading (arrow pointing down), significantly higher than that of the comparison school (arrow pointing up), or if there is no statistically significant difference between the two schools (circle). The blank diagonal represents where a school is compared against itself.

If there were no error around the overall scores for schools, Tables 5 through 9 would only consist of arrows pointing up and down, except for instances where two schools have the same overall score and are tied for rank. This is not the case. For example, in the business school rankings comparison table (Table 5) Harvard is listed first in the row and heading as it has the highest overall score among business schools. However, reading across the row, it appears that Harvard's overall score of 100 is not significantly different from that of nine other schools that are ranked beneath it. These include Stanford, which is tied for first rank with Harvard with an overall score of 100, and University of California, Berkeley, ranked tenth with a score of 90. Only schools ranked below tenth have scores that are significantly lower than Harvard's.

Tables 5 - 9
Statistical Significance of Comparisons of Overall Scores
in Five Areas (Data in the Form of Excell Spreadsheets)

- **Table 5: Business**

- **Table 6: Education**

- **Table 7: Law**

- **Table 8: Liberal Arts**

- **Table 9: National Universities**

In general, when the overall score for a school is compared to that of every other school in its ranking (top-fifty schools only), three groups emerge: schools that score significantly higher, schools that score significantly lower, and schools with scores that are not significantly different. This pattern is consistent across all the comparison tables. For example, among the business schools in Table 5, three distinct groupings emerge. The first group comprises 10 schools at the top of the rankings, extending from first-ranked Harvard to tenth-ranked University of California, Berkeley. These schools have scores that are not significantly different from each other but that are significantly higher than all other schools' scores. The second grouping extends from eleventh-ranked Dartmouth, University of California Los Angeles, and the University of Virginia to nineteenth-ranked Carnegie Mellon. These schools have scores that are not significantly different from each other but that are significantly lower than the top-ranked schools in the first group and significantly higher than the lower-ranked schools in the third grouping. The third group is the largest. It comprises 31 schools, extending from twentieth-ranked Indiana University to forty-eighth-ranked University of Georgia, University of Illinois-Urbana Champagne, and the University of Notre Dame. These schools all have scores that are not significantly different from each other but that are significantly lower than the scores of schools in the first two groups.

This three-groupings pattern is evident for all rankings except schools of education. There are only two groupings evident in Table 6. The first group comprises the top-three-ranked schools of education—Harvard University, Stanford University, and Teacher's College/Columbia University. These schools have scores that are not significantly different from each other but that are significantly higher than the scores for almost all other schools in the top fifty. The second group of schools extends from fourth-ranked University of California-Berkeley to the four schools tied for fiftieth rank. These schools all have scores that are not significantly different from each other but that are significantly lower than the scores of most schools in the top group. This two-grouping effect occurs because schools of education are more sensitive to changes in the indicators used than other types of schools. This results in larger standards errors around their overall score and fewer significant differences between the scores of neighboring schools.

Conclusions and Recommendations

The results of these analyses show that, given the number and annual nature of changes to each ranking formula, it is generally not possible to interpret year-to-year shifts in a school's rank in terms of change in its relative academic quality. Depending on the ranking, it is possible to make cross-year comparisons of a school's relative performance on between a third to three-quarters of the individual indicators used. While not experiencing much change to their ranking formula over time, schools of education have experienced markedly more movement in their ranks than other schools. It is not evident why this has occurred or what it says about the *U.S. News* rankings as a measure of the relative quality of these schools. The overall rate of change in the ranking formulas appears to be slowing and it was possible to make cross-year comparisons of schools' ranks for almost all rankings between 1999 and 2000.

The results of the error analyses call into question the use of overall scores to assign schools to individual ranks. The analyses show that when interpreting scores for school with the aid of their standard errors, precision blurs and schools start to group in bands rather than discrete ranks. The results confirm the critics' sense of unease at the precision of a single score, particularly in the case of the education rankings.

At least five recommendations can be made for improving the interpretability and usefulness of the *U.S. News* rankings.

1. First, *U.S. News* needs to stabilize their ranking methodology. This is particularly important since the rankings are annual in nature and imply some kind of comparability. A related issue to consider is whether the rankings need to be annual in nature. While there is an obvious commercial value to annual rankings, particularly one that keeps changing the winners, it is doubtful whether there is an educational or consumer value.

2. Second, *U.S. News* needs to recognize the uncertainty around schools' overall scores. The results of this analysis suggest that it would be more accurate to group schools in bands than to assign them discrete ranks. This approach would avoid the misleading effect that small changes in a school's rank from year to year produces in terms of the public perception's of its academic quality.

3. Third, the schools of education rankings need to be reassessed since they do not seem to "hold together." Better comparisons might emerge if they were divided into two more conceptually coherent groups (e.g., those that are primarily research oriented and those that are primarily teacher-training oriented.) *U.S. News* already does this for schools of medicine—i.e., there is an overall ranking of medical schools as well as a ranking of schools that focus on the training of primary-care physicians.

4. Fourth, in order to be accountable to consumers, *U.S. News* needs to make available all data used to create the rankings. Currently, *US News* only publishes information for the top-ranked schools and less or no information on lower-ranked schools. While space constraints may make it difficult to publish this information in the magazine, no such restrictions apply on the *US News* website.

5. A final general recommendation is that *U.S. News* should adopt a model similar to that used by *Consumer Reports* for reporting its quality ratings. *Consumer Reports* rates products, but does not allow the product manufacturers to use these ratings in their advertising. Similarly, *U.S. News* should not allow schools to use their ratings in their promotional materials or other advertising. This approach might relieve some of the tension and debate that currently surrounds the rankings and make their annual arrival on newsstands a less stressful event for the higher education community.

Notes

[1.] The term "rankings," as used in this artcles, refers to a list of schools or universities that are ordered according to their overall score on a formula created by *U.S. News*. Thus, the business rankings are a list of business schools ordered according to their overall score on a formula that *U.S. News* uses to rank graduate schools of business, and the national university rankings are a list of schools ordered according to their overall score on a formula that *U.S. News* uses to rank national universities. The year appended to a ranking is the calendar year in which it was released, i.e., the 2000 education rankings were published in the year 2000.

[2.] It is worth noting that several of these indicators—such as test scores, reputation, research expenditure, and faculty awards—have been used traditionally to measure quality (Hattendorf, 1993: Webster, 1986). The *U.S. News* rankings differ from most other rankings in that they assign weights to these indicators in order to combine them and produce a composite score.

[3.] The full text of this letter is available at:
http://www-portfolio.stanford.edu:8050/documents/president/961206gcfallow.html

[4.] No data was available for schools below the top-50 for most of the rankings.

[5.] *U.S. News* does not make available in its magazine or on its website all the data it uses to rank schools, nor is this information available on request. On average, each ranking is missing information on two or three indicators. This was not a problem for this analysis, since the available indicators, as indicated by the adjusted R Squared values, almost perfectly replicated the overall scores produced by *U.S. News*. Thus, very little information was lost.

[6.] While the "error estimate" obtained is not strictly a standard error, since the indicators are not randomly sampled, it may still be viewed as a general indication of the uncertainty around an overall score due to changes in the indicators used to compute that score. In addition, it is probably a conservative estimate of the uncertainty around scores as the indicators chosen by *U.S. News* tend to be highly correlated. A random sample from the population of indicators would probably be less highly correlated, which would result in larger standard errors around schools' overall scores.

[7.] Since there are, on average, 50 schools in each ranking, around 49 t-test comparisons were made for each school in the rankings. In order to control for the increased probability of a significant finding due to chance alone, a Bonferroni adjustment was applied.

[8.] For more information see http://www.usnews.com/usnews/edu/college/corank.htm

[9.] *U.S. News* uses a modification of the classification system developed by the Carnegie Foundation for the Advancement of Teaching in order to classify colleges and universities. The Carnegie system is a generally accepted classification system for higher education.

References **219**

Art and Science Group, (1995). Influence of *U.S. News* rankings on college choice, *StudentPOLL: Market Intelligence for Higher Education*, 1(1).

Crenshaw, A. B. (1999). Colleges by the numbers. *Washington Post*, August 29, p. H01.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Hattendorf, L. C. (ed.) (1993). *Educational rankings annual*. Detroit, MI: Gale Research Inc.

Levin, S. (1999). Ignore college ranking—become an educated consumer. Washington Parent Magazine. Available at: http://www.washingtonparent.com/articles/9712/rankings.htm

Machung, A. (1998). What's at stake: College rankings and the new media. *Change*, July/August, pp. 13-16.

McDonough, P. M., Antonio, A. L., Walpole, M., and Perez, L. X. (1998). College rankings: Democratized knowledge for whom? *Research in Higher Education*, 39(5), pp. 513-37.

Monks, J., and Ehrenberg, R. G. (1999). *The impact of U.S. News and World Report college rankings on admissions outcomes and pricing policies at selective private institutions*. Cambridge, MA: National Bureau of Economic Research.

Mufson, S. (1999). Rankings all-important to GWU. *Washington Post*, Sunday, March 14, P.A01.

Pellegrini, F. (1999). Those bouncing college rankings—a 101. *Time*, August 31.

Smetanka, M. J. (1998). Magazine's new ratings on colleges don't rank high at "U." *Star Tribune*, February 20, p. 7B.

Stecklow, S. (1995). Cheat sheets: Colleges inflate SATs and graduation rates in popular guidebooks: Schools say they must fib to US News and others to compete effectively: Moody's requires the truth. *The Wall Street Journal*, Wednesday, April 5, pp. 1, A8.

Thompson, J. J., and Morse, R. J. (1998). An explanation of the *U.S. News* rankings: Putting the numbers into context. *U.S. News and World Report America's Best Colleges 1998*, p. 66-68.

Webster, D. S. (1986). *Academic quality rankings of American colleges and universities*. Springfield, IL: Charles C. Thomas.

Wright, B. A. (1990-91). The rating game: How the media affect college admission. *The College Board Review*, No. 158, Winter, pp. 12-17, 31.

About the Author

Marguerite Clarke
Boston College
Phone: 617-552-0665
Fax: 617-552-8419
Email: clarkemd@bc.edu

Marguerite Clarke is an Assistant Professor of Research at Boston College. She has a Ph.D. in educational research, measurement, and evaluation as well as degrees in bilingual/multicultural and elementary education. Her research interests include policy and technical issues surrounding large-scale testing and accountability programs; the impact of testing on teaching and learning; and the relationship between test use and educational opportunity and access for different student populations.

Appendix A

Current *U.S. News* College and Graduate School Ranking Methodology

The current method that *U.S. News* uses to produce college rankings has three basic steps. (Note 8) First, colleges in the U.S. are placed into categories based on mission and region. (Note 9) Colleges within each category are ranked separately. Second, *U.S. News* collects data from each school on up to 16 separate indicators of what it believes reflects academic quality. As Table 10 indicates, each indicator is assigned a weight in the ranking formula that

220

reflects the judgement of *U.S. News* about which measures of quality matter most. Column 4 of Table 10 shows the weight that each indicator (shown in column 3 of Table 10) receives within its category and column 2 shows the weight this category receives in the overall ranking formula. For example, a school's acceptance rate is 15 percent of its Student Selectivity category score or rank, and the Student Selectivity category contributes 15 percent to a school's overall score and rank.

Indicators are standardized and then combined (using weights) to produce an overall score for each school. These scores are re-scaled. The top school is assigned a value of 100, and the other schools' weighted scores are calculated as a proportion of that top score. Final scores for each ranked school are rounded to the nearest whole number and ranked in descending order. *U.S. News* publishes the individual ranks of only the top schools; the remainder is grouped into tiers.

Table 10
***U.S. News* Indicators and Weights for the 2000 College Rankings[a]**

Ranking Category	Category Weight	Indicator	Indicator Weight
Academic Reputation	25%	Academic Reputation Survey	100%
Student Selectivity	15%	Acceptance Rate Yield High School Standing— Top 10% SAT/ACT Scores	15% 10% 35% 40%
Faculty Resources	20%	Faculty Compensation Faculty With Top Terminal Degree Percent Full-time Faculty Student/Faculty Ratio Class Size, 1-19 Students Class Size, 50+ Students	35% 15% 5% 5% 30% 10%
Retention Rate	20%	Average Graduation Rate Average Freshmen Retention Rate	80% 20%
Financial Resources	10%	Educational Expenditures Per Student	100%
Alumni Giving	5%	Alumni Giving Rate	100%
Graduation Rate Performance	5%	Graduation Rate Performance	100%

[a]These indicators and weights are for the national liberal arts and national university rankings only.

A similar methodology is employed for the graduate school rankings. *U.S. News* collects data from each program on indicators of what it believes reflect academic quality. Each indicator is assigned a weight based on *U.S. News'* judgment about which measures matter most. Data are standardized, and standardized scores are weighted, totaled, and re-scaled so that the top school receives 100; other schools receive a percentage of the top score. Schools are then ranked based on the score they receive.

The five major disciplines examined yearly are business, education, engineering, law, and medicine. Master's and doctoral programs in areas such as the arts, sciences, social sciences, humanities, library science, public affairs, and various health fields are ranked only by reputation and are generally evaluated every third year. The specific indicators and weights used for rankings within each of the five major disciplines are outlined in Tables 11 through 15.

Table 11
***U.S. News* Indicators and Weights for the 2000 Business Rankings**

Ranking Category	Category Weight	Indicator	Indicator Weight
Reputation	40%	Academic Survey Non-academic Survey	60% 40%
Placement Success	35%	Mean Starting Salary and Bonus Employment at Graduation and Three Months Later	40% 20% and 40%
Student Selectivity	25%	Mean Graduate Management Admission Test Scores Mean Undergraduate Grade Point Average Proportion of Applicants Accepted	65% 30% 5%

Table 12
U.S. News Indicators and Weights for the 2000 Education Rankings

Ranking Category	Category Weight	Indicator	Indicator Weight
Reputation	40%	Academic Survey Non-academic Survey	60% 40%
Student Selectivity	20%	Average Verbal, Analytical and Quantitative GREs Proportion of Applicants Accepted	30% each 10%
Faculty Resources	20%	Ratio of Full-time Doctoral and Master's Degree Candidates to Full-time Faculty Percent of Faculty Given Awards Number of Doctoral and Master's Degrees Granted in the past school year Proportion of Graduate Students Who Are Doctoral Candidates	25% and 20% 20% 15% and10% 10%
Research Activity	20%	Total Research Expenditures Research Expenditures Per Faculty Member	75% 25%

Table 13. U.S. News Indicators and Weights for the 2000 Engineering Rankings

Ranking Category	Category Weight	Indicator	Indicator Weight
Reputation	40%	Academic Survey Non-academic Survey	60% 40%
Student Selectivity	10%	Average Quantitative and Analytical GREs Proportion of Applicants Accepted	45% each 10%
Faculty Resources	25%	Ratio of Full-time Doctoral and Master's Degree Candidates to Full-time Faculty Proportion of Faculty Members of NAE Number of Ph.D Degrees Granted in the last school year Proportion of Faculty Holding Doctoral Degrees	25% and 10% 25% 20% 20%
Research Activity	25%	Total Research Expenditures Research Expenditures Per Faculty Member	60% 40%

Table 14
U.S. News Indicators and Weights for the 2000 Law Rankings

Ranking Category	Category Weight	Indicator	Indicator Weight
Reputation	40%	Academic Survey Non-academic Survey	60% 40%
Student Selectivity	25%	Median LSAT Scores Median Undergraduate GPA Proportion of Applicants Accepted	50% 40% 10%
Placement Success	20%	Employment Rates at Graduation and Nine Months Later Bar Passage Rate	30% and 60% 10%
Faculty Resources	15%	Average Expenditures Per Student For Instruction etc. Student to Teacher Ratio Average Expenditures Per Student For Financial Aid etc. Total Number of Volumes in Law Library	65% 20% 10% 5%

Table 15
U.S. News Indicators and Weights for the 2000 Medicine and Primary-Care (in parentheses where

Ranking Category	Category Weight	Indicator	Indicator Weight
Reputation	**40%**	Academic Survey Non-academic Survey	50% (60%) 50% (40%)
Student Selectivity	**20%**	Mean MCAT Scores Mean Undergraduate Grade Point Average Proportion of Applicants Accepted	65% 30% 5%
Faculty Resources	**10%**	Ratio of Full-time Science and Clinical Faculty to Full-time Students	100%
Primary Care Rate (Primary Care Only)	**30%**	The Percentage of MDs From a School Entering Primary-care Residencies, Averaged Over 1997, 1998, and 1999	100%
Research Activity (Medicine only)	**30%**	Total Dollar Amount of National Institutes of Health Research Grants Awarded to the Medical School and its Affiliated Hospitals, Averaged for 1998 and 1999	100%

Education Policy Analysis Archives

Are Increasing Test Scores in Texas Really a Myth, or is Haney's Myth a Myth?

Laurence A. Toenjes
University of Houston

A. Gary Dworkin
University of Houston

Abstract
Pass rates by Texas tenth-graders on the high school exit exam improved from 52 percent in 1994 to 72 percent in 1998. In his article "The Myth of the Texas Miracle in Education" (*EPAA*, August 2000) Professor Walt Haney argued that some part of this increased pass rate was, as he put it, an illusion. Haney contended that the combined effects of students dropping out of school prior to taking the 10th grade TAAS and special education exemptions accounted for much of the increase in TAAS pass rates. Relying on the same methodology and data that Haney used, we demonstrate that his conclusion is incorrect. None of the 20 percent improvement in the TAAS exit test pass rate between 1994 and 1998 is explained by combined increases in dropout rates or special education exemptions.

All may not be right with education in Texas. But neither is it all wrong, as Walter Haney would have everyone believe, judging by his article "The Myth of the Texas Miracle In Education."(Note 1) Haney wastes no space in getting to his main conclusion. In the first paragraph of the introduction he asserts that "In this article, I review evidence to show that the "miracle" of education reform in Texas is really a myth and illusion."(Note 2) However, he generously invites each reader to arrive at his or her own decision as to whether he was fair in arriving at his conclusions. "I leave it to others to judge how fair-minded I have been in recounting this version of the Texas miracle."(Note 3)

There is no attempt here to deal with all of the issues raised by Prof. Haney. In fact, only one issue is dealt with, but it is the one that is central to his thesis, namely whether or to what extent increases in dropout rates in Texas were caused by the Texas Assessment of Academic Skills (TAAS) exit test and the extent to which any increase in dropouts resulted in an unwarranted increase in the calculated pass rate on that test. At a minimum, this is a good example of how two different analysts can draw opposite conclusions from the same data.

Haney asserted at numerous points in his article, and elsewhere, that the TAAS exit test directly resulted in an

225

increase in dropouts, which in turn inflated apparent increases in pass rates on the exit test between the years 1994 to 1998.

Typical of statements attributed to Haney include the following: "I would guess at least half of the apparent increases are a mirage resulting from increasing numbers of students being excluded from test results—either because they dropped out of school or they've been misclassified as special education students.(Note 4) "The Texas miracle in education is a myth," said Walter Haney, a Boston College researcher who studies test statistics. Texas schools, he said, have some of the nation's highest dropout rates, and the system of accountability that Bush touts helps drive tens of thousands of students, mostly minorities, to quit school each year—a loss that in turn boosts test scores, he said. (Note 5)

The Haney article, as published in the *Education Policy Analysis Archives* is a distillation of his two-year effort as an expert—and presumably paid—witness for the Mexican American Legal Defense and Education Fund (MALDEF) in their suit against the State of Texas.(Note 6) The plaintiffs claimed that the exit test in Texas, first administered to students in the spring of their tenth grade, is unfair and discriminates against minority students. The goal of the suit was to prevent the State of Texas from continuing to make passing the exit test a condition for high school graduation. Much of Haney's effort was directed towards trying to convince Judge Prado that the 20-point increase in the pass rate on the exit test between 1994 and 1998 was due to substantial increases in numbers of students who dropped out before even taking the exit test. Haney tried to demonstrate that if students who would most likely fail the exam dropped out of school before they were scheduled to take the test in the spring of their tenth grade, then the calculated pass rate from the remaining students would be greater as a consequence. In addition, of course, the alleged increase in dropouts, especially if they occurred disproportionately among minority students, would directly demonstrate the damaging impact upon minority students.

Professor Haney went beyond factual arguments and attempted to impute motive to administrators and teachers. For example, "These results clearly support the hypothesis advanced in my December 1998 report, namely that after 1990 schools in Texas have increasingly been retaining students, disproportionately Black and Hispanic students, in grade nine **in order to make their grade 10 TAAS scores look better** [emphasis added] (Haney, 1998, pp. 17-18).(Note 7)

After having convinced himself, at least, that students are intentionally retained so that the exit test pass rate would be higher, he concluded: "Hence, it is fair to say that the soaring grade 10 TAAS pass rates are not just an illusion, but something of a fraud from an educational point of view."(Note 8)

While it was important to Haney and his clients to attempt to demonstrate that the very existence of the TAAS exit exam in some way, intentionally or not, caused an increase in grade nine retention and subsequent increase in dropouts, it was perhaps even more important to their case to demonstrate that these factors, in turn, were responsible for the dramatic improvement in TAAS exit test pass rates. For if the pass rate increase could be shown to be primarily a result of the increased retentions and dropouts (and also, perhaps, increases in the use of special education exemptions) then the primary justification for the TAAS-based accountability system itself would be discredited. That is, as long as the state could demonstrate that the academic performance of students who remained in school was improving, then it could be argued that this benefit offset if not outweighed the alleged increase in the number of dropouts. But if the improved performance, as measured by the pass rate on the exit test, was an illusion, as Haney asserted, due to the very increases in dropouts and special education exemptions, then the overall impact of the exit test could be argued to be a burden to the state as a whole as well as to the additional students who dropped out and thereby failed to earn their high school diplomas.

As pointed out by Haney, Judge Prado held that the hypothesis that "schools are retaining students in ninth grade in order to inflate tenth-grade TAAS results was not supported with legally sufficient evidence demonstrating the link between retention and TAAS (Prado, 2000, p. 27)"(Note 9) In fact, it will be shown below, utilizing data contained in the Haney article itself, that the possible impact of increased dropouts and ninth grade retentions actually *decreased* during the period 1994 to 1998, using Haney's own methodology. Judge Prado was correct.

In setting out to quantify the relationship between increases in ninth grade retentions, student dropouts and TAAS pass rates Haney relied upon the ratio over time of eleventh grade enrollment in a given year to sixth grade enrollment five years earlier. It is important that the reader be familiar with Haney's own justification for the use of this procedure, and to see how he misused it in reaching his conclusion. Following a discussion of the relationship of grade 9 to grade 8 enrollments, we find the following:

> At the same time, the analyses of progress for grade 6 cohorts presented in Section 5.3 revealed that grade 6 to grade 11 progression ratios for Whites and minorities varied by not more than 5% during the 1990s (for Whites, the ratio was consistently between 85% and 89%; and for minorities between 75% and 80%). The reason for focusing here on progress to grade 11 is because the data on enrollments is from the fall whereas TAAS is taken in the spring. But if students progress to grade 11, they presumably have taken the exit level version of TAAS in spring of the tenth grade.

> What this suggests is that the majority of the apparent 20-point gain in grade 10 TAAS pass rates cannot be attributed to exclusion of the types just reviewed. Specifically, if rates of progress from

226

grades 6 to grade 11 have varied by no more than 5% for cohorts of the classes of the 1990s, this suggests even if we take this as an upper bound, the extent to which increased retention and dropping out before fall grade 11, and add 2% for the increased rate of grade 10 special education classification, we still come up with less than half of the apparent 20-point gain in grade 10 TAAS pass rates between 1993 [sic, 1994] and 1998.(Note 10)

After actually looking at the data, Haney was forced to admit that "less than half of the apparent 20-point gain in grade 10 TAAS pass rates" could be accounted for. It is shown below that considerably less than half was accounted for by the data which he used.

To emphasize, if a student is enrolled in the fall of his or her eleventh grade level, presumably the student would have been in the tenth grade the previous spring of the same calendar year, and would therefore have been in the pool of students who would have taken the tenth grade exit test for the first time.(Note 11) Several issues are glossed over here, such as (a) students repeating grade 11, (b) students beginning public school in Texas as eleventh graders (immigrants or students previously enrolled outside the public school system), (c) students who may have taken the exit test in the spring but who dropped out over the following summer, and (d) changes in the proportion of students exempted from taking the exit TAAS. Regarding this latter, Haney presented data that suggested that the number of tenth grade students in special education had increased by approximately two percent during the 1994 to 1998 period. (Note 12) Accepting the simplifying assumptions in Haney's procedure, the results hinge on whether or not the ratio of eleventh graders to sixth graders (five years earlier) increased or decreased during the 1994 to 1998 period. Haney referred to these ratios as grade 6 to grade 11 progression ratios. For clarity, let the progression ratios for 1994 and 1998 be defined as follows:

Progression Ratio(94) = (Grade 11 enrollment in fall 1994/Grade 6 enrollment in fall 1989)

Progression Ratio(98) = (Grade 11 enrollment in fall 1998/Grade 6 enrollment in fall 1993)

If the data show that Progression Ratio(98) is less than Progression Ratio(94), then Haney has made his case. On the assumption that the 1998 ratio was reduced by an increased rate of dropout behavior occurring before the exit tests were taken in the spring of 1998, or an increase in ninth grade retentions, and assuming that all of the additional students who thereby did not take the exit test would have failed the test, then the pass rate would have increased by approximately the same rate as the increase in the rate of dropouts and ninth grade retentions, relative to the pass rate with no change in these phenomena.

Of course, the opposite might also occur, in which case an adjustment to the observed pass rate in the opposite direction should be made. That is, assume the proportion of dropouts decreased, or smaller proportions of students were retained in the ninth grade. This would cause an increase in Progression Ratio(98) as compared to Progression Ratio(94), Still assuming, with Haney, that all of the marginal students would fail the exit test, then the pass rate for the set of students in 1998 that would be comparable to the equivalent set in 1994 would be greater than the observed pass rate for all students tested in 1998. This is because, under such assumptions, greater proportions of low performing students would be tested in 1998 then in 1994. In short, if Haney's methodology would call for a negative adjustment to the pass rate if the progression ratio decreased from 1994 to 1998, then applying the same methodology to the opposite outcome, (i.e., an increase in the progression ratios) should require an upwards adjustment to the pass rate actually observed 1998.

What really happened? It is difficult to know, if one relies only upon Haney's text. In the first of the two paragraphs, there is a reference to progression ratios during the 1990s without mention of whether they tended upwards or downwards, stating only that "they varied by not more than 5% during the 1990s..." In the second paragraph quoted above, he then suggests that this vague 5% variation which occurred sometime "during the 1990s" can be taken as an upper bound of the impact of grade retentions and increased dropouts upon the exit test pass rate increase.

One cannot tell from Prof. Haney's own statements whether the ratio of grade 11 to grade 6 students increased or decreased during the particular period 1994 to 1998, which is the period during which the exit test pass rates increased by 20-points. In proposing to adjust the increase in the passing rates, what happened before or after this interval is irrelevant. Hence, reference to a range of variation "during the 1990s" is not helpful. Nor is reference to variation in these ratios for Whites and minorities taken separately. The 20-point improvement in test pass rates includes all students taking the exit test, in all subjects. If the grade 11/grade 6 enrollment ratios of Whites increased by 5%, but those for minorities decreased by 5%, they would approximately offset one another. If they both decreased by 5%, then the total effect would also be 5%. They are not additive. It is necessary to know what happened to the grade 11 to grade 6 enrollment ratios for *all* students in order that any change be relevant in adjusting changes to pass rates for all students. It is also necessary that they be based on grade 11 enrollments in 1994 and in 1998.

Fortunately, Haney included in an appendix the data necessary to clarify these ambiguities.(Note 13) The enrollment data, by grade level, are available there for the three major ethnic groups from 1989 to 1998. The three major ethnic groups included 98 percent of total enrollment in 1989. The major group not included was Asian-Americans, for which dropout behavior and TAAS performance is not a major issue. Their omission does not alter the results presented below in Figure 1 or in Table 1.

Ratio of Enrollments, Grade 11 to Grade 6

Grade 6 Enrollments Five Years Prior to Grade 11 Enrollments
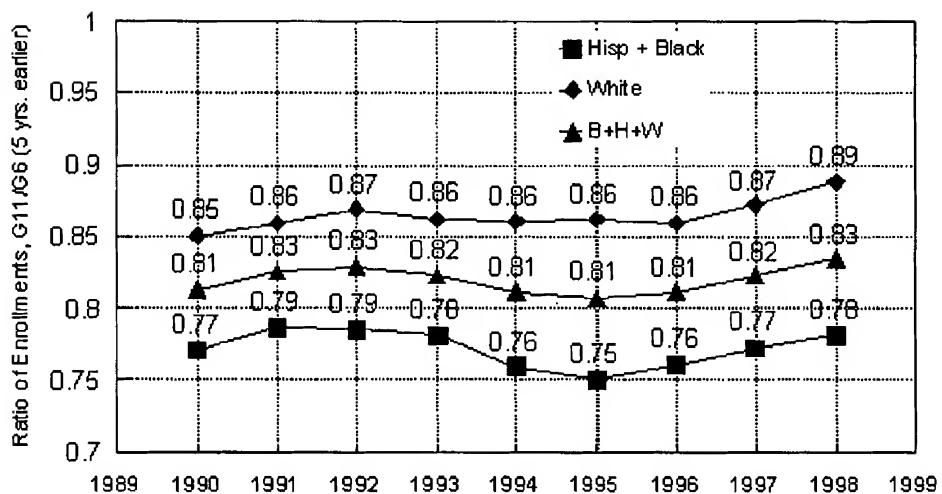
Data Source: Haney (2000), Appendix 7

Enrollment counts taken in fall of year. Ratios plotted on year corresponding to grade 11.

Figure 1. Progression Ratios, 1990–1998.

The progression ratios for grade 6 to grade 11 are presented in Figure 1. They are shown separately for African Americans and Hispanics combined, for non-Hispanic Whites, and for the totals of all three major ethnic groups. The ratios were calculated by dividing the appropriate eleventh grade enrollments by corresponding sixth grade enrollments five years earlier. The time scale in Figure 1 indicates the year for which the eleventh grade enrollments correspond. As pointed out by Haney (see above) the calendar year for which eleventh grade enrollments are recorded in the fall is also the year in which the same cohort of students would have been first administered the exit test the preceding spring. Therefore, a significant change in the progression ratio for a given year would be expected to have influenced the percentage of students passing the exit test in the same calendar year.(Note 14)

It is interesting to compare the data plotted in Figure 1 with Prof. Haney's comments. As he stated (above) the "progression ratios for Whites and minorities varied by not more than 5% during the 1990s..."(fn) Indeed, the highest ratio for minorities was 0.79, the lowest was 0.75 for a range of 0.04 or 4 percentage points. The high for Whites was 0.89 and the low was 0.85, for a range also of 4 percentage points. But it is important to note how these ratios changed over the specific interval of 1994 to 1998, which corresponds to the interval over which the 20-point increase in exit test pass rates occurred. It is not sufficient to merely note the "range". What is important is whether the ratios increase or decreased over this particular period. Additionally, it is the change in the ratio for all ethnic groups combined that is relevant to an attempt to explain the 20-point improvement in the test pass rate, as this improvement reflected the improved performance for all groups combined.

So what actually happened to the relevant progression ratios? As can be seen in Figure 1, the progression ratio for minorities *increased* from 0.76 in 1994 to 0.78 in 1998, albeit with a one percent decrease in 1995. For Whites, the progression ratio also *increased,* from 0.86 in 1994 to 0.89 in 1998. For all three major ethnic groups combined, the progression ratio increased from 0.81 in 1994 to 0.83 in 1998. Therefore, instead of adjusting the improvement in the pass rate from 20-points down to 15-points, as implied by Haney, it should be adjusted upwards by 2-points to 22-points due to the increase in the grade 6 to grade 11 progression ratio.

Granting a negative adjustment of 2 points due to an increase in special education exemptions among tenth-graders, the effect upon the improvement in the exit test pass rate attributable to special education exemptions, ninth-grade retentions, and dropout rates net to zero—no impact whatsoever.

Total enrollment figures for the three major ethnic groups for the relevant years are shown in Table 1. Once again, the results shown in column 5 of Table 1 are exactly opposite to the assertions by Haney. Instead of a negative adjustment of 5 points to the 20-point improvement in the exit test passing rate over the 1994 to 1998 period, a positive 2 point adjustment should be made, following the same logic.

Table 1
Texas Enrollments for African-American, Hispanic, and

Non-Hispanic White Students Grade 6 (1989-1993), Grade 11 (1994-1998), and Progression Ratios

Year (t-5)	Grade 6 Enrollments	Year (t)	Grade 11 Enrollments	G11/G6 Progression Ratio
(1)	(2)	(3)	(4)	(5)=(4)/(2)
1989	245,828	1994	199,379	0.811
1990	256,551	1995	207,140	0.807
1991	269,839	1996	218,822	0.811
1992	275,779	1997	226,794	0.822
1993	278,663	1998	232,441	0.834
4-year chg	+32,835		+33,062	

Data source: Haney (2000), Appendix 7, pages 138-139.

Notice also that grade 6 enrollment for the three major ethnic groups increased by 32,835 students between 1989 and 1993, while eleventh grade enrollment corresponding to the same cohorts, five years later, grew by 33,062 students from 1994 to 1998. The fact that grade 11 enrollments increased more than grade 6 enrollments certainly does not support Haney's claim of increasing dropouts and/or rates of ninth grade retentions during this time period.

Therefore, using Haney's own suggested methodology, and data which he himself provided, none of the improvement in the TAAS exit test pass rate has been shown to be a myth or otherwise fraudulently obtained. Instead of demonstrating that "at least half of the apparent increases are a mirage resulting from increasing numbers of students being excluded from test results", as he had claimed, his data and procedures account for none of the exit test pass rate improvement.

Prof. Haney invited readers of his article to form their own judgments as to the fairness of his analysis. We have formed our own judgment as to whether Professor Haney was fair-minded in his use of data and in the conclusions he drew from those data regarding the quality and effectiveness of education reform in Texas over the last decade. Like Professor Haney, we invite readers to arrive at their own judgments on this matter.

Notes

[1] Walter Haney, "The Myth of the Texas Miracle In Education," *Education Policy Analysis Archives*, Volume 8, Number 41, August 19, 2000, available at (http://cpaa.asu.edu/epaa/v8n41/). There are no page numbers on the article as posted on the website. The page numbers given below are to a printed version of that document.

[2] Haney., p. 7.

[3] Haney., p. 8.

[4] Debra Viadero, "Testing System in Texas Yet To Get Final Grade", *Education Week*, May 31, 2000, available at (http://educationweek.org/ew/ew_printstory.cfm?slug=38taas.h19).

[5] John Mintz, " 'Texas Miracle' Doubted: An Education 'Miracle,' or Mirage?," The Washington Post, April 21, 2000, p. A01.

[6] Haney, p. 7.

[7] Haney, p. 46. The reference to Haney, 1998 was to Haney, W. (1998)."Preliminary report on Texas Assessment of Academic Skills Exit Test (TAAS-X)." Chestnut Hill, MA: Boston College Center for the Study of Testing Evaluation and Educational Policy.

[8] Haney, p. 54.

[9] Haney, p. 46. The reference to Prado, 2000, was to Prado, E. (2000). Order in case of GI Forum Image De Tejas v.

Texas Education Western District of Texas (Civil Action No. SA-97-CA-1278-EP. Filed January 7, 2000. (GI Forum Image De Tejas V. Texas Education Agency, 87 F. Supp. 667 (W.D. Tex. 2000).

[10]Haney, p. 56.

[11]The exit exam includes tests in reading, mathematics, and writing. Each of these may be take more than once. The exit test pass rates being considered are based upon the first taking of these tests by (usually) tenth grade students.

[12]The increase in tenth grade students which were classified as receiving special education services is not disputed. However, whether or not the increase in special education classifications were appropriate or not is another matter.

[13]Haney, Appendix 7, p. 138-141.

[14]To the extent that students dropped out after taking the exit test in the tenth grade and therefore were not included in the next fall's enrollment counts, the grade6 to grade 11 progression ratio would overstate the likely impact on test scores.

About the Authors

Laurence A. Toenjes is Research Associate Professor of Sociology and Co-Founder of the Sociology of Education Research Group (SERG) at the University of Houston. He holds the Ph.D. degree in Economics from Southern Illinois University. A substantial portion of his research is in the area of school finance, and he has developed school finance models for several states.

A. Gary Dworkin is Professor of Sociology and Co-Founder of the Sociology of Education Research Group (SERG) at the University of Houston. He is past chair of the Department of Sociology at the University of Houston. He holds the Ph.D. in sociology from Northwestern University. The majority of his publications is in the areas of the sociology of education and race, ethnic, and gender relations and include nine books and numerous articles and book chapters. He is editor of a book series entitled "The New Inequalities," published by the State University of New York Press. During the first half of 2001 he served as Senior Research Fellow in Sociology at the Australian National University in Canberra ACT, Australia.

High-Stakes Testing, Uncertainty, and Student Learning

Audrey L. Amrein
Arizona State University

David C. Berliner
Arizona State University

Abstract

A brief history of high-stakes testing is followed by an analysis of eighteen states with severe consequences attached to their testing programs. These 18 states were examined to see if their high-stakes testing programs were affecting student learning, the intended outcome of high-stakes testing policies promoted throughout the nation. Scores on the individual tests that states use were not analyzed for evidence of learning. Such scores are easily manipulated through test-preparation programs, narrow curricula focus, exclusion of certain students, and so forth. Student learning was measured by means of additional tests covering some of the same domain as each state's own high-stakes test. The question asked was whether transfer to these domains occurs as a function of a state's high-stakes testing program.

Four separate standardized and commonly used tests that overlap the same domain as state tests were examined: the ACT, SAT, NAEP and AP tests. Archival time series were used to examine the effects of each state's high-stakes testing program on each of these different measures of transfer. If scores on the transfer measures went up as a function of a state's imposition of a high-stakes test we considered that evidence of student learning in the domain and support for the belief that the state's high-stakes testing policy was promoting transfer, as intended.

The uncertainty principle is used to interpret these data. That principle states "The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor." Analyses of these data reveal that if the intended goal of high-stakes testing policy is to increase student learning, then that policy is not working. While a state's high-stakes test may show increased scores, there is little support in these data that such increases are anything but the result of test preparation and/or the exclusion of students from the testing process. These distortions, we argue, are predicted by the uncertainty principle. The success of a high-stakes testing policy is whether it affects student learning, not whether it can increase student scores on a particular test. If student learning is not affected, the validity of a state's test is in question.

Evidence from this study of 18 states with high-stakes tests is that in all but one analysis, student learning is indeterminate, remains at the same level it was before the policy was implemented, or actually goes down when high-stakes testing policies are instituted. Because

clear evidence for increased student learning is not found, and because there are numerous reports of unintended consequences associated with high-stakes testing policies (increased drop-out rates, teachers' and schools' cheating on exams, teachers' defection from the profession, all predicted by the uncertainly principle), it is concluded that there is need for debate and transformation of current high-stakes testing policies.

This is an era of strong support for public policies that use high-stakes tests to change the behavior of teachers and students in desirable ways. But the use of high-stakes tests is not new, and their effects are not always desirable. "Stakes," or the consequences associated with test results, have long been a part of the American scene. For example, early in the 20th century, scores on the recently invented standardized tests could, for immigrants, result in entrance to or rejection from the United States of America. In the public schools test scores could uncover talent, providing entrance into programs for the gifted, or as easily, provide evidence of deficiencies, leading to placement in vocational tracks or even in homes for the mentally inferior. Test scores could also mean the difference between acceptance into, or rejection from, the military. And throughout early twentieth century society, standardized test scores were used to confirm the superiority or inferiority of various races, ethnic groups, and social classes. Used in this way, the consequences of standardized tests insured maintenance of the status quo along those racial, ethnic and class lines. So, for about a century, significant consequences have been attached to scores on standardized tests.

A Recent History of high-stakes Testing

In recent decades, test scores have come to dominate the discourse about schools and their accomplishments. Families now make important decisions, such as where to live, based on the scores from these tests. This occurs because real estate agents use school test scores to rate neighborhood quality and this affects property values. (Note 1) Test scores have been shown to affect housing prices, resulting in a difference of about $9,000 between homes in grade "A" or grade "B" neighborhoods. (Note 2) At the national and state levels, test scores are now commonly used to evaluate programs and allocate educational resources. Millions of dollars now hinge on the tested performance of students in educational and social programs.

Our current state of faith in and reliance on tests has roots in the launch of Sputnik in 1957. Our (then) economic and political rival, the Soviet Union, beat the United States to space, causing our journalists and politicians to question American education with extra vigor. At that time, state and federal politicians became more actively engaged in the conduct of education, including advocacy for the increased use of tests to assess school learning. (Note 3)

The belief that the achievement of students in U.S. schools was falling behind other countries led politicians in the 1970s to instigate a minimum competency testing movement to reform our schools. (Note 4) States began to rely on tests of basic skills to ensure, in theory, that all students would learn at least the minimum needed to be a productive citizen.

One of these states was Florida. After some hasty policy decisions, Florida implemented a statewide minimum competency test that students were required to pass prior to being graduated. Florida's early gains were used as an example of how standards and accountability systems could improve education. However, when perceived gains hit a plateau and differential pass rates and increased dropout rates among ethnic minorities and students from low socioeconomic backgrounds were discovered, Florida's testing policy was postponed. (Note 5)

In the 1980s, the minimum competency test movement was almost entirely discarded. Beyond what was happening in Florida, suggestions that minimum competency tests promoted low standards also raised concerns. In many schools the content of these tests became the maximum in which students, particularly in urban schools, became competent. (Note 6) It was widely perceived that minimum competency tests were "dumbing down" the content learned in schools.

In 1983, the National Commission on Education released *A Nation at Risk,* (Note 7) the most influential report on education of the past few decades. *A Nation at Risk* called for an end to the minimum competency testing movement and the beginning of a high-stakes testing movement that would raise the nation's standards of achievement drastically. Although history has not found the report to be accurate, (Note 8) it argued persuasively that schools in the United States were performing poorly in comparison to other countries and that the United States was in jeopardy of losing its global standing. Citing losses in national and international student test scores, deterioration in school quality, a "diluted" and "diffused" curriculum, and setbacks on other indicators of U.S. superiority, the National Commission on Education triggered a nationwide panic regarding the weakening condition of the American education system.

Despite its lack of scholarly credibility *A Nation at Risk* produced massive effects. The National Commission on

Education called for more rigorous standards and accountability mechanisms to bring the United States out of its purported educational recession. The Commission recommended that states institute high standards to homogenize and improve curricula and rigorous assessments be conducted to hold schools accountable for meeting those standards. The Commission and those it influenced intended to increase what students learn in schools. This report is an investigation of how well that explicitly intended outcome of high-stakes testing programs was achieved. We ask, below, whether increases in school learning are actually associated with increases in the use of high-stakes tests? Although it appears to be a simple question, it is very difficult to answer.

The Effects of *A Nation at Risk* on Testing in America

As a result of *A Nation at Risk*, state policymakers in every state but Iowa developed educational standards and every state but Nebraska implemented assessment policies to check those standards. (Note 9) In many states high-stakes, or serious consequences, were attached to tests in order to hold schools, administrators, teachers, and students accountable for meeting the newly imposed high standards.

In fixing high-stakes to assessments, policymakers borrowed principles from the business sector and attached incentives to learning and sanctions to poor performance on tests. High performing schools would be rewarded. Under performing schools would be penalized, and to avoid further penalties, would improve themselves. Accordingly, students would be motivated to learn, school personnel would be forced to do their jobs, and the condition of education would inevitably improve, without much effort and without too great a cost per state. What made sense, in theory, gained widespread attention and eventually increased in popularity as a method for school reform.

Arguments in Support of High-stakes Tests.

At various times over the past years different arguments have been used to promote high-stakes tests. A summary of these follows:

- students and teachers need high-stakes tests to know what is important to learn and to teach;
- teachers need to be held accountable through high-stakes tests to motivate them to teach better, particularly to push the laziest ones to work harder;
- students work harder and learn more when they have to take high-stakes tests;
- students will be motivated to do their best and score well on high-stakes tests; and that
- scoring well on the test will lead to feelings of success, while doing poorly on such tests will lead to increased effort to learn.

Supporters of high-stakes testing also assume that the tests:

- are good measures of the curricula that is taught to students in our schools;
- provide a kind of "level playing field," an equal opportunity for all students to demonstrate their knowledge; and that
- high-stakes tests are good measures of an individual's performance, little affected by differences in students' motivation, emotionality, language, and social status.

Finally, the supporters believe that:

- teachers use test results to help provide better instruction for individual students;
- administrators use the test results to improve student learning and design better professional development for teachers; and that
- parents understand high-stakes tests and how to interpret their children's scores.

The validity of these statements in support of high-stakes tests have been examined through both quantitative and qualitative research, and by the commentary of teachers who work in high-stakes testing environments. A reasonable conclusion from this extensive corpus of work is that these statements are true only some of the time, or for only a modest percent of the individuals who were studied. The research suggests, therefore, that *all* of these statements are likely to be false a good deal of the time. And in fact, some research studies show exactly the opposite of the effects anticipated by supporters of high-stakes testing. (Note 10)

The Heisenberg Uncertainty Principle Applied to the Social Sciences

For many years the research and policy community has accepted a social science version of Heisenberg's Uncertainty Principle. That principle is *The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor.* (Note 11) When applied to a high-stakes testing environment, this principle warns us that attaching serious personal and educational consequences to performance on tests for schools, administrators, teachers, and students, may have distorting and corrupting effects. The distortions and corruptions that accompany high-stakes tests make inferences about the

meanings of the scores on those tests uncertain. If there is uncertainty about the meaning of a test score the test may not be valid. Unaware of this ominous warning, supporters of high-stakes testing, particularly politicians, have caused high-stakes testing to proliferate. The spread of high-stakes tests throughout the nation is described next.

Current High-stakes Testing Practices

Today, twenty-two states offer schools incentives for high or improved test scores. (Note 12) Twenty states distribute financial rewards to successful schools, and nineteen states distribute financial rewards to improved schools.

Punishments are attached to school scores twice as often as rewards, however. Forty-five states hold schools accountable for test scores by publishing school or district report cards. Twenty-seven of those states hold schools accountable through rating and ranking mechanisms; fourteen have the power to close, reconstitute, or take over low performing schools; sixteen have the authority to replace teachers or administrators; and eleven have the authority to revoke a school's accreditation. In low performing schools, low scores also bring about embarrassment and public ridicule.

For administrators, threats of termination and cuts in pay exist, as does the potential for personal bonuses. In Oakland, California, for example, city school administrators can receive a 9% increase in pay for good school performance with a potential for an additional 3% increase—1% per increase in reading, math and language arts. (Note 13)

For teachers, low average class scores may prevent teachers from receiving salary increases, may influence tenure decisions, and in sixteen states may be cause for dismissal. Only Texas has linked teacher evaluations to student or school test results, but more states have plans to do so in the future.

High average class scores may also bring about financial bonuses or raises in pay. Eleven states disperse money directly to administrators or teachers in the most improved schools. For example, California recently released each school's Academic Performance Index (API). This is based almost entirely on Stanford 9 test scores. Schools showing the biggest gains were to share $677 million in rewards while low performing schools in which personnel did not raise student achievement scores were to face punishments. (Note 14) In addition, teachers and administrators in 1,346 California schools that demonstrated the greatest improvements over the past 2 years were to share $100 million in bonus rewards, called Certificated Staff Performance Incentive Bonuses, ranging from $5,000 to $25,000 per teacher. Although over $550 million had already been disbursed to California schools, the distribution of the staff bonuses was deferred because some teachers who posted gains on the API scale, but felt they were denied their share of the reward money, filed a lawsuit against the state. (Note 15) The court found in favor of the state.

Schools and teachers were not the only targets of rewards and punishments for test performance. Policy makers also attached serious consequences to performance on tests for individual students.

Although test scores are often promoted as diagnostic tools useful for identifying a student's achievement deficits and assets, they are rarely used for such purposes when they emanate from large-scale testing programs. Two major problems are the cause of this. First, test scores are often reported in the summer after students exit each grade and second, there are usually too few items on any one topic or area to be used in a diagnostic way. (Note 16) As a result of these factors, scores on large-scale assessments are most often used simply to distribute rewards and sanctions. This contributes to the corruptions and distortions predicted by the social science version of Heisenberg's Uncertainty Principle.

The special case of scholarships

The distortions and corruptions predicted by the Uncertainty Principle find fertile ground for developing when high scores on a test result in special diplomas or scholarships. Attaching scholarships to high performance on state tests is a relatively new concept, yet six states have already begun granting college scholarships and dispersing funds to students with distinguished performance on tests. (Note 17) Michigan is a perfect example of the corruptions and distortions that occur when stakes are high for a quantitative social indicator.

The Michigan imbroglio. In spring 2000, Michigan implemented its Merit Award Scholarship program in which 42,700 students who performed well on the Michigan Educational Assessment Program high school tests were rewarded with scholarships of $2,500 or $1,000 to help pay for in-state or out-of-state college tuition, respectively. (Note 18)

There is quite a story behind these scholarships, however. (Note 19) In 1996, Michigan became the 13th state to sue the nation's leading cigarette manufacturers to recover health care costs encumbered by the state to treat smoking-related diseases developed by Michigan's poor and disadvantaged citizens. The care and treatment of these citizens placed a financial burden on the states, so they sued the tobacco companies for financial compensation. Michigan won approximately $384 million to recover some of these health care costs and then decided to distribute approximately 75% of this money among high school seniors with high test scores as Merit Award Scholarships. The remainder of the money went to health related needs and research, more or less unrelated to smoking or disease treatment. Thus,

235

the monies that were awarded to the state did not go to the victims at the center of the lawsuit—Michigan's poor and indigent suffering from tobacco related diseases—but went instead to those students who scored the highest on the Michigan Educational Assessment Program high school test. These were Michigan's relatively wealthier students who had the highest probability of enrolling in college even without these scholarships. (Note 20)

Approximately 80% of the test-takers in an affluent Michigan neighborhood earned scholarships while only 6% of the test-takers in Detroit earned scholarships. (Note 21) One in three white, one in fourteen African American, one in five Hispanic, and one in five Native American test takers received scholarships. (Note 22) In addition, from 1982 to 1997, while education spending for needy students increased 193%, education spending for merit based programs such as the merit scholarships increased by 457% in Michigan. (Note 23) Tests have often been defended because they can distribute or redistribute resources based on notions of "merit." But too often the testing programs become thinly disquised methods to maintain the status quo and insure that funds stay in the hands of those who need them least.

Michigan is now being sued by a coalition that includes students, the American Civil Liberties Union of Michigan (ACLU), the Mexican American Legal Defense and Education Fund (MALDEF), and the National Association for the Advancement of Colored people (NAACP). They are arguing that Michigan is denying students scholarships based on test scores that are highly related to race, ethnicity, and educational advantages. Michigan appears to be a state where high-stakes testing has had a corrupting influence.

The satisfying effects of punishing the slackers. Connecting high-stakes tests with rewards for high performance, such as in the example above, is not nearly as prevalent as have been punishments attached to student scores that are judged to be too low. Punishments are used three times as often as rewards. Policy makers appear to derive satisfaction from the creation of public policies that punish those they perceive to be slackers.

Throughout the nation low scores are used to retain students in grade, using the slogan of ending "social promotion." Promotion or retention is already contingent on test performance in Louisiana, New Mexico, and North Carolina, while four more states have plans to link promotion to test scores in the next few years. (Note 24)

Low scores may also prevent high school students from graduating from high school. Whether a student passes or fails high school graduation exams – exams that purportedly test a high school student's level of knowledge in core high school subjects – is increasingly being used as the *only* determinant of whether some students graduate or whether students are entitled to a regular high school diploma or merely a certificate of attendance.

In fact, high school graduation exams are the assessments with the highest, most visible, and most controversial stakes yet. When *A Nation at Risk* was released, only three states (Note 25) had implemented high school graduation exams, then referred to as minimum competency tests on which students' *basic* skills were tested. But in *A Nation at Risk,* the commission called for more rigorous examinations on which high school students would be required to demonstrate mastery in order to receive high school diplomas. (Note 26) Since then, states have implemented high school graduation exam policies with greater frequency.

Now, almost two decades later, eighteen states (Note 27) have developed and employed high school graduation exams and nine more states (Note 28) have high-school graduation exams underway. The frequency with which high school graduation exams have become key components of states' high-stakes testing policies has escalated almost linearly over the past twenty-three years and will continue to do so for at least the next six years (see Figure 1).
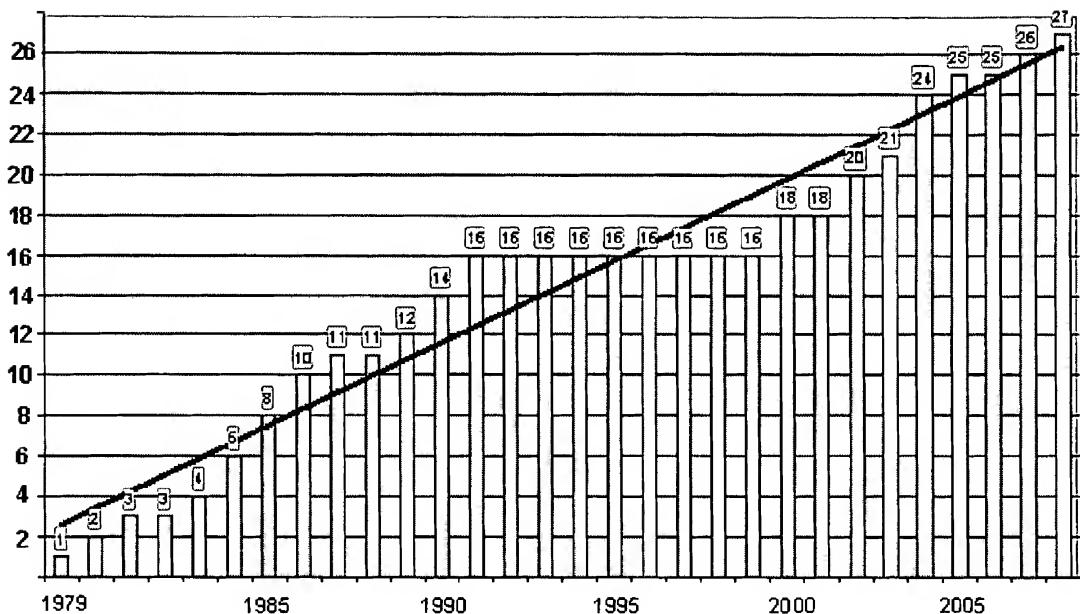
Figure 1. Number of states with high school graduation exams 1979–2008 (Note 29)

Who Uses high-stakes Tests?

Analyses of these data reveal that high school graduation exams are:

- more common in states that allocate less money than the national average per pupil for schooling as compared to the nation. High school graduation exams are found in around 60% of the states in which yearly per pupil expenditures are lower and in about 45% of the states in which yearly per pupil expenditures are higher than the national average. (Note 30)

- more likely to be found in states that have more centralized governments, rather than those with more powerful county or city governments. Of the states that have more centralized governments, 62% have or have plans to implement high school graduation exams. Of the states that have less centralized governments, only 37% have or have plans to implement high school graduation exams. (Note 31)

- more likely to be found in the highly populated states and states with the largest population growth as compared to the nation. (Note 32) For example, 76% of the country's most highly populated states and only 32% of the country's smallest states have or have plans to implement high school graduation exams. Looking at growth, not just population we find that 76% of the states with the greatest population growth and only 32% of the states with the lowest population growth from 1990–2000 have or have plans to implement high school graduation exams. (Note 33)

- most likely to be found in the Southwest and the South. High school graduation exams are currently in use in 50% of the southwestern states and 66% of the southern states. Analyses also suggest that high school graduation exams will become even more common in these regions in the future. By the year 2008, high school graduation exams will be found in 75% of the southwestern and southern states.

High school graduation exams will probably continue to be randomly dispersed throughout 50% of the states in the Northeast and least likely to be found in 33% of the mid-western states. The western states, over the next decade, will have the greatest increase in the number of states with high school graduation exams by region. While 10% percent of the western states have already implemented high school graduation exam policies, 50% of these states will have implemented these exams by the year 2008. (Note 34)

More important for understanding high-stakes testing policy is that high school graduation exams are more likely found in states with higher percentages of African Americans and Hispanics and lower percentages of Caucasians as compared to the nation. Census Bureau population statistics helped to verify this. (Note 35) Seventy-five percent of the states with a higher percentage of African Americans than the nation have high school graduation exams. By 2008 81% of such states will have implemented high school graduation exams. Sixty-seven percent of the states with a higher percentage of Hispanics than the nation have high school graduation exams. By 2008 89% of such states will have implemented high school graduation exams. Conversely, 13% of the states with a higher percentage of Caucasians than the nation have implemented high school graduation exams. By 2008 29% of such states will have

implemented high school graduation exams. *In other words, high school graduation exams affect students from racial minority backgrounds in greater proportions than they do white students.* If these high-stakes tests are discovered not to have their intended effects, that is, if they do not promote the kinds of transfer of learning and education the nation desires, the mistake will have greater consequences for America's children of color.

Similarly, high school graduation exams disproportionately affect students from lower socioeconomic backgrounds. High school graduation exams are more likely to be found in states with the greatest degrees of poverty as compared to the nation. Economically disadvantaged students are most often found in the South and the Southwest and least often found in the Northeast and Midwest. As noted, states in the South and the Southwest are most likely to have high-stakes testing policies. Further, 69% of the states with child poverty levels greater than the nation have or have plans to implement high school graduation exams. Seventy percent of the states with the greatest 1990–1998 increases in the number of children living in poverty have or have plans to implement such exams. (Note 36) That is, *high school graduation exams are more likely to be implemented in states that have lower levels of achievement, and the always present correlate of low achievement, poorer students.* Again, if these high-stakes tests are discovered not to have their intended effects, that is, if they fail to promote transfer of learning and education in its broadest sense, as the nation desires, the mistake will have greater consequences for America's poorest children.

Matters of national standards and implementation of high-stakes tests are less likely to be of concern for the reform of relatively elite schools, (Note 37) that are more often found in regions other than the South and Southwest. Perhaps this helps to explain the more extensive presence of high-stakes tests in the South and Southwest. This seems a reasonable hypothesis especially when one purpose of high-stakes testing is to raise student achievement levels in educational environments perceived to be failing.

It should be noted, however, that there is considerable variability in these data. All states with high rates of children in poverty have not adopted high-stakes testing policies while some states with lower rates of children in poverty have. In states with higher or lower levels of poverty, however, schools that exist within poor rural and urban environments are still more frequently targeted by these policies. Although legislators promote these policies, claiming high standards and accountability for all, schools that already perform well on tests are not the targets for these policies; poor, urban, under performing schools are. But, for different reasons, support for high-stakes testing receives support in both high and low achieving school districts. In successful schools and districts, high-stakes testing policies are acceptable because the scores on those tests merely confirm the expectations of the community. Thus, in successful communities, the tests pose little threat and also have little incentive value. (Note 38) In poorer performing schools high-stakes testing policies often enjoy popular support because, it is thought, at the very least, that these tests will raise standards in a state's worst schools. (Note 39)

But if high-stakes testing policies do not promote learning, that is, if they do not appear to be leading to education in the most profound sense of that term, then the tests will not turn out to have any use in successful communities and schools, nor will they improve the schools attended by poor children and ethnic minorities. If, in addition, the tests have unintended consequences such as narrowing the curriculum taught, increasing drop out rates and contributing to higher rates of retention in grade, they would not be good for any community. But these unintended negative consequences would have a greater impact on the families and neighborhoods of poor and minority students.

Faith in testing. The effects of high-stakes tests on students is well worth pursuing since it is unquestionably a "bull market" for testing. (Note 40) The faith state legislators have put into tests, albeit blind, has increased dramatically over the past twenty years. (Note 41) The United States tests its children more than any other industrialized nation, has done so for well over thirty years, (Note 42) and will continue to depend on even more tests as it attempts to improve its schools. At the national level, President Bush has been unquestionably successful in passing his "No Child Left Behind" plan that calls for even more testing – annual high-stakes testing of every child in the United States in grades 3 through 8 in math and reading. Republicans and Democrats alike have endorsed high-stakes testing policies for the nation making this President Bush's only educational proposal that has claimed bipartisan support. (Note 43) According to the President and other proponents, annual testing of every child and the attachment of penalties and rewards to their performance on those tests, will unequivocally reform education. Despite the optimism, the jury is still out on this issue.

Many researchers, teachers and social critics contend that high-stakes testing policies have worsened the quality of our schools and have created negative effects that severely outweigh the few, if any, positive benefits associated with high-stakes testing policies. Because testing programs and their effects change all the time, reinterpretations of the research that bears on this issue will be needed every few years. But at this time, in contradiction to all the rhetoric, the research informs us that states that have implemented high-stakes testing policies have fared worse on independent measures of academic achievement than have states with no or low stakes testing programs. (Note 44) The research also informs us that high-stakes testing policies have had a disproportionate negative impact on students from racial minority and low socioeconomic backgrounds. (Note 45)

In Arizona, for example, officials reported that in 1999 students in poor and high-minority school districts scored lower than middle-class and wealthy students on Arizona's high-stakes high school graduation test, the AIMS (Arizona's Instrument to Measure Standards). Ninety-seven percent of African Americans, Hispanics, and Native Americans failed the math section of the AIMS, a significantly greater proportion of failures than occurred in the white community, whose students also failed the test in great numbers. (Note 46) Due to the high failure rates for

different groups of students, as well as various psychometric problems, this test had to be postponed.

In Louisiana parents requested that the office for civil rights investigate why nearly half the children in school districts with the greatest numbers of poor and minority children had failed Louisiana's test, after taking it for a second time. (Note 47) In Texas, in 1997, only one out of every two African American, Mexican American, and economically disadvantaged sophomores passed each section of Texas' high-stakes test the TAAS – Texas' Assessment of Academic Skills. In contrast, four out of every five white sophomores passed. (Note 48) In Georgia, two out of every three low-income students failed the math, English, and reading sections of Georgia's competency tests. *No* students from well-to-do counties failed any of the tests and more than half exceeded standards. (Note 49)

The pattern of failing scores in these states are quite similar to the failure rates in other states with high school graduation exams and are illustrative of the achievement gap between wealthy, mostly white school districts and poor, mostly minority school districts. (Note 50) It appears that a major cause of these gaps is that high-stakes standardized tests may be testing poor students on material they have not had a sufficient opportunity to learn.

Education, Learning, and Training: Three Goals of Schooling

In this report we look at just one of the distorting and corrupting possibilities suggested by Heisenberg's Uncertainty Principle applied to the testing movement, namely, that *training* rather than *learning* or general *education* is taking place in communities that rely on high-stakes tests to reform their schools. As will be become clearer, if we have doubt about the meaning of a test score, we must be skeptical about the validity of the test.

Our interest in these distinctions between training, learning and education stems from the many anecdotes and research reports we read that document the narrowing of the curriculum and the inordinate amount of time spent in drill as a form of test preparation, wherever high-stakes tests are used. The former president of the American Association of School Administrators, speaking also as the Superintendent of one of the highest achieving school districts in America, notes that:

> The issue of teaching to these tests has become a major concern to parents and educators. A real danger exists in that the test will become the curriculum and that instruction will be narrow and focused on facts.

> ... Teachers believe they spend an inordinate amount of time on drills leading to the memorization of facts rather than spending time on problem solving and the development of critical and analytical thinking skills. Teachers at the grade levels at which the test is given are particularly vulnerable to the pressure of teaching to the test.

> Rather than a push for higher standards, [Virginia's high-stakes] tests may be driving the system toward mediocrity. The classroom adaptations of "Trivial Pursuit" and "Do You Want to be a Millionaire?" may well result in higher scores on these standardized tests, but will students have acquired the breadth and knowledge to do well on other quality benchmarks, such as the SAT and Advanced Placement exams? (Note 51)

This is our concern as well. Any narrowing of the curriculum, along with the confusion of training to pass a test with broader notions of learning and education are especially problematic side effects of high-stakes testing for low-income students. The poor, more than their advantaged peers, need not only the skills that training provides but need the more important benefits of learning and education that allow for full economic and social integration in our society.

To understand the design of this study and to defend the measures used for our inquiry requires a clarification of the distinctions between the related concepts of *education*, *learning* (particularly *school learning* and the concept of *transfer of learning*), and *training*. For most citizens it is education (the broadest and most difficult to define of the concepts) that is the goal of schooling. Learning is the process through which education is achieved. But merely demonstrating acquisition of some factual or procedural knowledge is not the primary goal of school learning. That is merely a proximal goal.

The proper goal of school learning is both more distal and more difficult to assess. The proper goal of school learning is transfer of learning, that is, the application or use of what is learned in one domain or context to that of another domain or context. School learning in the service of education focuses deliberately on the goal of broad (or far) transfer. School instruction that can be characterized as training is ordinarily a narrow form of learning, where transfer of learning is measured on tasks that are highly similar to those used in the training. Broad or far measures of transfer, the appropriate goal of school learning, are different from the measures typically used to assess the outcomes of training.

More concretely, training in holding a pencil, or of doing two-column addition with regrouping, or memorizing the names of the presidents, is expected to yield just that. After training to do those things is completed students should be

able to write in pencil, add columns of numbers, and name the presidents. The assessments used to measure their newly acquired knowledge are simple and direct. On the other hand, learning to write descriptive paragraphs, arguing about how numbers can be decomposed, and engaging in civic activities should result in better writing, mathematics and citizenship. To inquire whether that is indeed the case, much broader and more distal measures of transfer are required and these kinds of outcomes of education are much harder to measure.

Although enormously difficult to define, almost all citizens agree that school learning is designed to produce an "educated" person. Howard Gardner provides one voice for these aspirations by claiming that students become educated by probing, in sufficient depth, a relatively small set of examples from the disciplines. In Gardner's curriculum teachers lead students to think and act in the manner of scientists, mathematicians, artists, or historians. Gardner advocates deep and serious study of a limited set of subject matter to provide students with opportunities to deal seriously with the genuine and profound ideas of humankind.

> I believe that three very important concerns should animate education; these concerns have names and histories that extend far back into the past. There is the realm of *truth*—and its underside, what is false or indeterminable. There is the realm of *beauty* — and its absence in experiences or objects that are ugly or kitschy. And there is the realm of *morality* — what we consider to be good, and what we consider to be evil. (Note 52)

Gardner's "educated" student thinks like those in the disciplines because the students learn the forms of argument and proof that are appropriate to a discipline. Thus tutored, students are able to analyze the fundamental ideas and problems that all humans struggle with. It is a discussion and project-oriented curriculum, with minimum concern for test preparation as a separate activity. Gardener's discipline-based curriculum is explicitly concerned with transfer to a wide array of human endeavors. Despite the difficulty in obtaining evidence of this kind of transfer of learning, there is ample support for this kind of curriculum. Earl Shorris recently demonstrated the effect of this kind of curriculum with desperately poor people who were given the chance to study the disciplines with excellent and caring teachers. (Note 53) The experience of studying art, music, moral philosophy, logic, and so forth, transformed the lives of these impoverished young adults.

Minnesota Senator Paul Wellstone also understands that school learning is not an end in itself. For him, our educational system should be designed to produce an "educated" person, someone for whom transfer of what is learned in school is possible:

> Education is, among other things, a process of shaping the moral imagination, character, skills and intellect of our children, of inviting them into the great conversation of our moral, cultural and intellectual life, and of giving them the resources to prepare to fully participate in the life of the nation and of the world." (Note 54)

Senator Wellstone, however, sees a problem with this goal:

> Today in education there is a threat afoot,...: the threat of high-stakes testing being grossly abused in the name of greater accountability, and almost always to the serious detriment of our children." (Note 55)

The Senator, like many others, recognizes the possible distorting and corrupting effects of high-stakes testing. He worries about compromising the education of our students, because of "a growing set of classroom practices in which test-prep activities are usurping a substantive curriculum." (Note 56) The Senator is concerned that test preparation for for the assessment of narrow curricular goals will turn out to be more like training than like the kind of learning that promotes transfer. And if that were to be the case, the test instruments themselves are likely to be narrow and near measures of transfer, as befits training programs. If this scenario were to occur, then broad and far measures of transfer, the indicators, we hope, of the educated person that we hold as our ideal, might not become part of the ways in which we assess what is being learned in our schools.

To reiterate: education (in some broad and hard-to-define way) is our goal. School learning is the means to accomplish that goal. But, as a recent National Academy of Seience/National Research Council report on school learning makes clear, schooling that too closely resembles training, as in preparation for testing, *cannot* accomplish the task the nation has set for itself, namely, the development of adaptive and educated citizens for this new millennium. (Note 57) Of course, school learning that promotes transfer is only a necessary, and not a sufficient condition, to bring forth an educated person. The issue, however, is whether high-stakes tests, with their potential for distorting and corrupting classroom life, can overcome the difficulties inherent in such systems, and thereby bring about the transformation in student achievements sought by all concerned with public education. One of the nation's leading experts on measurement has thought about this issue:

> As someone who has spent his entire career doing research, writing, and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing that the major uses of tests for student and school accountability during the past 50 years have improved education and student learning in dramatic ways.

Unfortunately, I cannot. Instead, I am led to conclude that in most cases the instruments and technology have not been up to the demands that have been placed on them by high-stakes accountability. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high-stakes are attached to them. The unintended negative effects of high-stakes accountability uses often outweigh the intended positive effects." (Note 58)

Transfer of learning and test validity. This report looks at one of the effects claimed for high-stakes testing: that states with high-stakes tests will show evidence that some kind of broad learning, rather than just some kind of narrow training, has taken place. It is well known that test preparation, meticulous alignment of the curriculum with the test, as well as rewards and sanctions for students and other school personnel, will almost always result in gains on whatever instrument is used by the state to assess its schools. Scores on almost all assessment instruments are quite likely to go up as school administrators and teachers *train* students to do well on tests such as the all-purpose widely-used SAT-9s in California, or the customized Texas Assessment of Academic Skills (TAAS), the Arizona Instrument to Measure Standards (AIMS), or the Massachusetts Comprehensive Assessment System (MCAS). We ask a more important question than "Do scores rise on the high-stakes tests?" We ask whether there is evidence of student learning, *beyond the training that prepared them for the tests they take,* in those states that depend on high-stakes tests to improve student achievement? We seek to know whether we are getting closer to the ideal we all hold of a broadly educated student, or whether we are instead developing students that are much more narrowly trained to be good test takers. It is important to note that this is not just a question of how well the nation is reaching its intended outcomes, it is also an equally important psychometric question about the validity of the tests, as well.

The National Research Council cautions that "An assessment should provide representative coverage of the content and processes of the domain being tested, so that the score is a valid measure of the student's knowledge of the broader [domain], not just the particular sample of items on the test." (Note 59)

So the score a student obtains on a high-stakes test must be an indicator of transfer or generalizability or that test is not valid. The problem is that:

1. tests almost always are made up of fewer items than the number actually needed to thoroughly assess the entire domain that is of interest;
2. testing time, as interminable as it may seem to the students, is rarely enough to adequately sample all that is to be learned from a domain; and
3. teachers may narrow what is taught in the domain so that the scores on the test will be higher, though by doing this, the scores are then invalid since they no longer reflect what the student knows of the entire domain.

These three factors work against having high-stakes test scores accurately reflect students' domain scores in areas such as reading, writing, science, etc. Because of this constant threat of invalidity, attaching high-stakes to achievement tests of this type may be impossible to do sensibly. (Note 60)

How might this show up in practice? Unfortunately there is already research evidence that reading and writing scores in Texas may not reflect the domains that are really of interest to us. The Heisenberg Uncertainty Principle applied to assessment seems may be at work distorting and corrupting the Texas system. The ensuing uncertainty about the meaning of the test scores in Texas requires skepticism about whether that state obtained valid indicators of the domain scores that are really of interest. That is, we have no assurance that the performance on the test indicates what it is supposed to, namely, transfer or generalizability of the performance assessed to the domain that is of interest to us. For example,

> ... high school teachers report that although practice tests and classroom drills have raised the rate of passing for the reading section of the TAAS at their school, many of their students are unable to use those same skills for actual reading. These students are passing the TAAS reading section by being able to select among answers given. But they are not able to read assignments, to make meaning of literature, to complete reading assignments outside of class, or to connect reading assignments to other parts of the course such as discussion and writing.

> Middle school teachers report that the TAAS emphasis on reading short passages, then selecting answers to questions based on those short passages, has made it very difficult for students to handle a sustained reading assignment. After children spend several years in classes where "reading" assignments were increasingly TAAS practice materials, the middle school teachers in more than one district reported that [students] were unable to read a novel even two years below grade level. (Note 61)

A similar phenomenon exists in testing writing, where a single writing format is taught—the five paragraph persuasive essay. Each paragraph has exactly five sentences: a topic sentence, three supporting sentences, and a concluding sentence much like the introductory sentence. The teachers call this "TAAS writing," as opposed to "real writing."

Teachers of writing who work with their students on developing ideas, on finding their voice as writers, and on organizing papers in ways appropriate to both the ideas and the papers' intended audience find themselves in conflict with this prescriptive format. The format subordinates ideas to form, sets a single form out as "the essay," and produces predictably, rote writing. Writing as it relates to thinking, to language development and fluency, to understanding one's audience, to enriching one's vocabulary, and to developing ideas has been replaced by TAAS writing to this format. (Note 62)

California also has well documented instances of this. The curriculum was so narrowed to reflect the high-stakes SAT 9 exam, and the teachers under such pressure to teach just what is on the test, that they voluntarily felt obliged to add a half hour a day of unpaid teaching time to the school schedule. As one teacher said:

> This year [we] ... extended our day a half hour more. And this is exclusively to do science and social studies. ... We think it's very important for our students to learn other subjects besides Open Court and math ... because in upper grades, their literature, all that is based on social studies, and science and things like that. And if they don't get that base from the beginning [in] 1st [and] 2nd grade, they're going to have a very hard time understanding the literature in upper grades .... There is no room for social studies, science. So that's when we decided to extend our day a half hour .... But this is a time for us. With that half hour, we can teach whatever we want, and especially in social studies and science and stuff, and not have to worry about, "OK, this is what we have to do." It's our own time, and we pick what we want to do. (Interview, 2/19/01) (Note 63)

In this school the stress to teach to the test is so great that some teachers violate their contract and take an hourly cut in pay in order to teach as their professional ethics demand of them. Such action by these teachers—in the face of serious opposition by some of their colleagues— is a potent indicator of how great the pressure in California is to narrow the curriculum and relentlessly prepare students for the high-stakes test. The paradox is, that by doing these things, the teachers actually invalidate the very tests on which they work so hard to do well. It is not often pointed out that *the harder teachers work to directly prepare students for a high-stakes test, the less likely the test will be valid for the purposes it was intended.*

Test preparation associated with high-stakes testing becomes a source of invalidity if students had differential test preparation—as often happens in the case of rich and poor students who take the SAT for college entrance. But even if all the students had intensive test preparation the potential for invalidity exists because the scores on the test may then no longer represent the broader domain of knowledge for which the test score was supposed to be an indicator. Under either of these circumstances, where there is differential preparation for the tests by *different* groups of students, or intensive test preparation by *all* the students, there is still a way to make a distinction between training effects and the broader more desirable learning effects. That distinction can be made by using transfer measures, that is, other measures of the same domain as the high-stakes test but where no intensive test preparation occurred. The scores of students on tests of the same or similar domains as those measured by the high-stakes test can help to answer the question about whether learning in the broad domain of knowledge is taking place, as intended, or whether a narrow form of learning is all that occurs from the test preparation activities. If scores on these other tests rise along with the scores on the state tests then genuine learning would appear to be taking place. The claim that transfer within the domain is occurring can then be defended, and support will have been garnered for the high-stakes testing programs now sweeping the country. We will now examine data that help to answer these questions about whether broad-based learning or narrow forms of training are occurring.

Design of the Study

The purpose of this study is to inquire whether the high-stakes testing programs promote the transfer of learning that they are intended to foster. A second report in this series inquires if there have been negative side-effects of high-stakes testing for economically disadvantaged and ethnic minority students (see "The Unintended Consequences of high-stakes Testing by A. L. Amrein & D. C. Berliner, forthcoming, at http://www.edpolicyreports.org/). The sample of states used to assess the intended and unintended effects of high-stakes testing are the eighteen states that have the most severe consequences, that is, the highest stakes associated with their K–12 testing policies: Alabama, Florida, Georgia, Indiana, Louisiana, Maryland, Minnesota, Mississippi, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, South Carolina, Tennessee, Texas, and Virginia. Table 1 describes the stakes that exist in each of these states at this time.

Table 1
Consequences/"Stakes" in K–12 Testing Policies in States that
Have Developed Tests with the Highest Stakes (Note 64)

States	Total Stakes	Grad. exam[a]	Grade prom. exam[b]	Public report cards[c]	Id. low perform.[d]	$ awards to schools[e]	$ awards to staff[f]	State may close low perform.[g]	State may replace staff[h]	Students may enroll else-where[i]	$ awards to students[j]
Alabama	6	X		X	X	X		X	X		

BEST COPY AVAILABLE

State		a	b	c	d	e	f	g	h	i	j
Florida	6	X		X	X	X	X			X	
Georgia	5	X	2004 (Note 65)	X	X	X	X	2004			
Indiana	6	X		X	X	X		X		X	
Louisiana	7	X	X (Note 66)	X	X			X	X	X	
Maryland	6	X		X	X	X		X	X		
Minnesota	2	X		X							
Mississippi	3	X		X	X	2003		2003			
Nevada	6	X		X	X			X	X		X
New Jersey	4	X		X	X	X					
New Mexico	7	X	X (Note 67)	X	X	X		X	X		
New York	5	X		X	X			X	X		
North Carolina	8	X	X (Note 68)	X	X	X	X	X	X (Note 69)		
Ohio	6	X	2002 (Note 70)	X	X	X	X				X
South Carolina	6	X	2002 (Note 71)	X	X	X		X	X		
Tennessee	6	X		X	X	X	X	X			
Texas	8	X	2003 (Note 72)	X	X	X	X	X	X (Note 73)	X	
Virginia	4	X		X	X			X			

aGraduation contingent on high school grad. exam.

bGrade promotion contingent on exam.

cState publishes annual school or district report cards.

dState rates or identifies low performing schools according to whether they meet state standards or improve each year.

eMonetary awards given to high performing or improving schools.

fMonetary awards can be used for "staff" bonuses.

gState has the authority to close, reconstitute, revoke a school's accred. or takeover low performing schools.

hState has the authority to replace school personnel due to low test scores.

iState permits students in failing schools to enroll elsewhere.

jMonetary awards or scholarships for in- or out of state college tuition are given to high performing students.

These states have not only the most severe consequences written into their K–12 testing policies but lead the nation in incidences of school closures, school interventions, state takeovers, teacher/administrator dismissals, etc., and this has occurred, at least in part, because of low test scores. (Note 74) Further, these states have the most stringent K–8 promotion/retention policies and high school graduation exam policies. They are the only states in which students are being retained in grade because of failing state tests and in which high school students are being denied regular high school diplomas, or are simply not graduating, because they have not passed the state's high school graduation exam. These data on denial of high school diplomas are presented in Table 2.

Table 2
Rates at Which Students Did Not Graduate or Receive a High School Diploma Due to Failing the State High School Graduation Exam (Note 75)

State (Note 76)	Grade in which students first take the exam	Percent of students who did not graduate or receive a regular high school diploma because they did not meet the graduation requirement (Note 77)	Year
Alabama*	10	5.5%	2001
Florida*	11	5.5%	1999
Georgia*	11	12%	2001
Indiana*	10	2%	2000

Louisiana	10 & 11	4%	2001
Maryland	6	4%	2000
Minnesota	8	2%	2001
Mississippi*	11	n/a (Note 78)	n/a
Nevada	11	3%	2001
New Jersey	11	6%	2001
New Mexico*	10	n/a	n/a
New York	n/a (Note 79)	10%	2000
North Carolina*	9 (Note 80)	7%	2000
Ohio	8	2%	2000
South Carolina	10	8%	1999
Tennessee	9	2.5%	2001
Texas	10	2%	2001
Virginia*	6	0.5%	2001

The effects of high-stakes tests on *learning* were measured by examining indicators of student learning, academic accomplishment and achievement *other* than the tests associated with high-stakes. These other indicators of student learning serve as the transfer measures that can answer our question about whether high-stakes tests show merely training effects, or show transfer of learning effects, as well. The four different measures we used to assess transfer in each of the states with the highest stakes were:

1. the ACT, administered by the American College Testing program;
2. the SAT, the Scholastic Achievement Test, administered by the College Board;
3. the NAEP, the National Assessment of Educational Progress, under the direction of the National Center for Education Statistics and the National Assessment Governing Board; and
4. the AP exams, the Advanced Placement examination scores, administered by the College Board.

In each state, for each test, participation rates in the testing programs were also examined since these vary from state-to-state and influence the interpretation of the scores a state might attain.

Transfer measures to assess the effects of high-stakes tests. As noted above, psychometricians teach us that one facet of validity is that the scores on a test are indicators of performance in the domain from which the test items are drawn. Thus, the score a student gets on a ten-item test of algebra, or on their driving test, ought to provide information about how that student would score on any of the millions of problems we could have chosen from the domain of algebra, or on how that student might drive in innumerable traffic situations. The score on the short classroom assessment, or on the test of driving performance, is actually an indicator of the students' ability to transfer what they have demonstrated that they have learned to the other items and traffic situations that are similar to those on the assessment. In a sense, then, we don't really care much about the score that was obtained on either test. What we really want to know is whether that student can do algebra problems or drive well in traffic. So we are interested in the score on the tests the student actually took only in so far as those scores represent what they know or can do in the domain in which we are interested. This study seeks to clarify the relationship between the score obtained on a high-stakes test and the domain knowledge that the test score represents.

If, as in some states, scores on the state test go up, it is proper to ask whether the scores are also going up on other measures of the same domain. That is precisely what a gain score on a state assessment should mean. Gain scores should be the indicators of increased competency in the domain that is assessed by the tests, and that is why transfer measures that assess the same domain are needed. (Note 81)

If the high-stakes testing of students really induces teachers to upgrade curricula and instruction or leads students to study harder or better, then scores should also increase on other independent assessments. (Note 82) So we used the ACT, SAT, NAEP and AP exams as the other independent assessments, as measures of transfer. We are not alone in using these four measures to assess transfer of learning. For example, one analyst of the Texas high-stakes program believes: "If Texas-style systemic reform is working as advertised, then the robust achievement gains that TAAS reports should also be showing up on other achievement tests such as the National Assessment of Educational Progress (NAEP), Advanced Placement exams and tests for college admission." (Note 83)

In addition, the RAND Corporation recently used this same logic to investigate the validity of impressive gains on Kentucky's high-stakes tests. The researchers compared the students' performance on Kentucky's state test with their performance on comparable tests such as the NAEP and the ACT. Gains on the state test did not match gains on the NAEP or ACT tests. They concluded the Kentucky state test scores were seemingly inflated and were not a meaningful indicator of increased student learning in Kentucky. (Note 84)

In assessing the effects of testing in Texas, other RAND researchers noted "Evidence regarding the validity of score gains on the TAAS can be obtained by investigating the degree to which these gains are also present on other measures of these same skills." (Note 85)

Because some test data from the states with high-stakes tests do not show evidence of learning on some of the transfer measures, journalist Peter Schrag noted that "...the unimpressive scores on other tests raise unavoidable questions about what the numbers really mean [on the high-stakes tests] and about the cost of their achievement." (Note 86)

The National Research Council also supports transfer measures of the type we use by relying on such data in their own analysis. They note, with dismay, that "There is some evidence to indicate that improved scores on one test may not actually carry over when a new test of the same knowledge and skills is introduced." (Note 87)

Sampling concerns. In each state the ACT and SAT tests are designed to measure the achievements of various percentages of the 60–70 percent of the total high school students in a state who intend to go to college. Within each state these tests probably attract a broad sample of students intending to go to college, while the AP tests are probably given to a more restricted and higher achieving sample of students. But in all three cases the samples are *not* representative of the state's high school graduates. However, these are all high-stakes tests for the students, with each test influencing their future. Thus, their motivation to do well on the state's high-stakes test and these other indicators of achievement is likely to be similar. This leads to a conservative test of transfer of learning, because it ought to be easier to find indicators of transfer, if it occurs, among these generally higher ability, more motivated students, rather than in a sample that included all the students in a state.

Motivation to achieve well may be diminished in the case of the NAEP because no stakes are attached to those tests. But the NAEP state data is obtained from a random sample of the states' schools, and thus may provide the most representative sample among the four measures of transfer of learning we use. Nevertheless, even with NAEP there is a problem. At each randomly selected school it is the local school personnel who decide if individual students will participate in NAEP testing. As will become clear later, sometimes the participation rates in NAEP testing seem suspect, leading to concerns about the appropriateness of the NAEP sample, as well.

In each high-stakes state, from the year in which the first graduating class was required to pass a high school graduation examination, we asked: What happened to achievement in the domains assessed by the American College Test (ACT), in the domains assessed by the Scholastic Achievement Test (SAT), in the domains assessed by the National Assessment of Educational Progress (NAEP), (Note 88) and in the domains assessed by the Advanced Placement (AP) tests. We asked also how participation rates in these testing programs changed and might have affected interpretations of any effects found.

An archival time-series research design was chosen to examine the state-by-state and year-to year data on each transfer measure. Time-series studies are particularly suited for determining the degree to which large-scale social or governmental policies make an impact. (Note 89) In archival time-series designs strings of observations of the variables of interest are made before, and after, some policy is introduced. The effects of the policy, if any, are apparent in the rise and fall of scores on the variable of interest.

We may consider the implementation of the state policy to engage in high-stakes testing as the independent variable, or treatment, and the scores from year to year on the ACT, SAT, NAEP and AP tests, before and after the implementation of high-stakes testing, as four dependent variables of interest. Relationships between the treatments and effects (between independent and dependent variables) are demonstrated by studying the pattern in the trend lines before and after the intervention(s), that is, before and after it was mandatory to pass state tests. (Note 90) Table 3 presents the dates at which high school graduation requirements of this type were first introduced in the eighteen states under study.

Table 3
Years in Which High School Graduation Exams
Affected Each Graduating Class (Note 91)

State	Year in which the state's 1st graduation exam policy was introduced	Graduating classes required to pass different graduation exams to receive a regular high school diploma.				
		1st Exam Class of...	2nd Exam Class of...	3rd Exam Class of...	4th Exam Class of...	5th Exam Class of...
Alabama	1983	1985	1993	2001, 2002, 2003		

Florida	1976	1979	1990	1996	2003	
Georgia	1981	1984	1995	1997, 1998	Future (Note 92)	
Indiana	1996	2000				
Louisiana	1989	1991	2003, 2004			
Maryland	1981	1987	2007			
Minnesota	1996	2000				
Mississippi	1988	1989	2003, 2004, 2005, 20061			
Nevada	1979	1981	1985	1992	1999	2003
New Jersey	1981	1984	1987	1995	2003, 2004, 2006	
New Mexico	1988	1990				
New York	1960s (Note 93)	1985	1995	2000, 2001, 2002, 2003, 2004, 2005		
North Carolina	1977	1980	1998 (Note 94)	2005		
Ohio	1991	1991	1994	2007		
South Carolina	1986	1990	2005, 2006, 2007	Future		
Tennessee	1982	1986	1998	2005		
Texas	1980	1983 (Note 95)	1987	1992	2005	
Virginia	1983	1986	2004			

BEST COPY AVAILABLE

Two strategies were used to help evaluate the strength of the effects of the high-stakes testing policy, and our confidence in those effects. First, data points before the introduction of the tests provided baseline information. (Note 96) Whether changes in the transfer measure occurred was determined by comparing the post intervention data with the baseline or pre-intervention data. If there was a change in the trend line for the data, just after intervention occurred, it was concluded that the treatment had an effect.

Secondly, national trend lines were positioned alongside state trend lines to help control for normal fluctuations and extraneous influences on the data. (Note 97) The national group was used as a nonequivalent comparison group to help estimate how the dependent variable would have oscillated if there had been no treatment. (Note 98) The national trend lines controlled for whether effects at the state level were genuine or just reflections of national trends. Figure 2, using actual data from the state of Alabama, and presented again in Appendix B, illustrates how the archival time series and our analyses of effects worked.
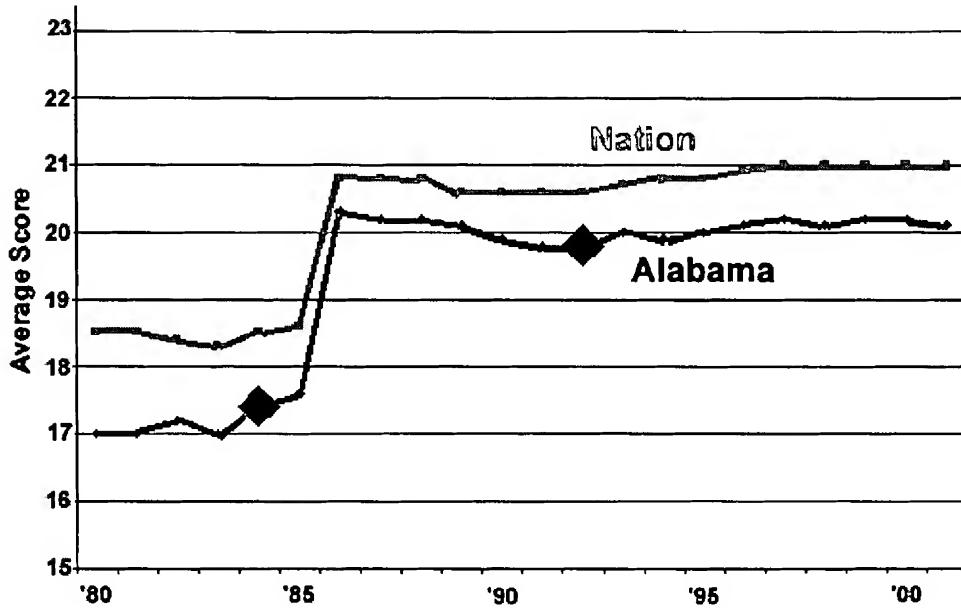
Figure 2. (From Appendix B) Analysis of the American College Test (ACT), Alabama (Note 99)

Alabama implemented its 1st high school graduation exam in 1983. It was a prerequisite for graduation that first affected the class of 1985. Alabama's 2nd exam first affected the class of 1993. The enlarged diamond shape signifies the year *before* the 1st graduating class was required to pass the exam. The policy intervention occurs in the year following the large diamond. From these data we conclude that:

- From 1984–1985 Alabama gained .1 point on the nation.
- From 1984–1992 Alabama gained .3 points on the nation.
- From 1992–1993 Alabama gained .1 point on the nation.
- From 1992–2001 Alabama lost .1 point to the nation.

To interpret these data, one inspects the state trend line and notes from the bold diamond shapes that there were two different points at which Alabama instituted high-stakes tests. After the first test was implemented, there was a score gain on the ACT in Alabama. (Note 100) After the second test there was an equally modest rise in Alabama's ACT scores. But in each case the national trend line showed similar effects, which moderates those conclusions. We can conclude from plotting the ACT scores each year that: 1) there were, indeed, small short term gains on the ACT in the year after new high-stakes tests were instituted; and 2) that the long term effects that may have occurred were substantial after the first test, but resulted in a small negative effect after the second high-stakes test was instituted. As can be seen, the national trend lines are quite important for interpreting the effects of a high-stakes testing policy on a measure of transfer.

A combined national trend line was used because the creation of a comparison group from the 32 states with no or low stakes attached to their tests was not feasible. Designation of which category a state was in changed from year to year so there were never clear cases of "states with high-stakes tests" and a comparison group made up of "states without high-stakes tests" across the years. Using the combined national trend line was the best comparison group available, even though this trend line included each of the states that were under analysis and the other 17 states that we designated as high-stakes states and were also the object of study. Because of these factors there are some difficulties in the comparison of the state and national trend lines, perhaps introducing some bias toward or against finding learning effects when comparing state trend lines with the national trend lines. If such bias exists, we believe its effects would be minimal.

Sources of Data

In an archival time series analysis, effects of the independent variable were measured using historical records and data collected from agency and governmental archives (Note 101) and extensive telephone calls and emails to and from agency personnel and directors. The following state-level data archives were collected:

American College Test (ACT)

- ACT composite scores – 1980–2001

247

- ACT participation rates – 1994–2001

SAT

- SAT composite scores – 1977–2001
- SAT participation rates – 1991–2001

National Assessment of Educational Progress (NAEP)

- NAEP Grade 4 Mathematics composite scores – 1992, 1996, 2000
- NAEP Grade 8 Mathematics composite scores – 1990, 1992, 1996, 2000
- NAEP Grade 4 Reading composite scores – 1992, 1994, 1998
- NAEP Grade 8 Reading composite scores – 1998

Advance Placement (AP)

- Percentage of 11th / 12th graders who took AP exams 1991–2000
- Percentage of 11th / 12th graders receiving a 3 or above 1995–2000

State summaries for each of the 18 states with the highest stakes written into their K–12 testing policies were constructed to facilitate the time series analysis. These are presented in Appendix A. The summaries include contextual and historical information about each state's testing policies. Each summary should help readers gain more insight about each state's testing policies and the values each state attributes to high-stakes tests, beyond the information offered in Table 1. Most importantly, each summary includes background information regarding the key intervention points, or years in which graduating seniors were first required to pass different versions of high school graduation exams as summarized in Table 3. These intervention points were illustrated in each archival time series graph, and each interpretation of state data relied on what happened after these key points in time. The archival time series graphs for each of the transfer measures we used are included in the different Appendices. The data associated with each of the transfer measures will now be described.

The American College Test (ACT) and the
Scholastic Achievement Test (SAT)

The American College Test (ACT) (Note 102) and Scholastic Achievement Test (SAT) (Note 103) are the two predominant entrance exams taken by students prior to enrolling in higher education. College-bound students take the ACT or SAT to meet in-state or out-of-state university enrollment requirements. Scores on these tests are used by college admissions officers as indicators of ability and academic achievement and are used in decisions about whether an applicant has the minimum level of knowledge to enter into, and prosper, at the college to which they applied. Although many studies have been conducted questioning the usefulness of these tests in predicting a student's actual success after enrolling in college, they continue to be widely used by universities when accepting students into their institutions.. (Note 104).

Despite questions about their predictive validity, both ACT and SAT scores can be considered as sensible indicators of academic achievement in the domains that constitute the general high school curriculum of the United States. Averaged at the state level, both tests can be thought of as external and alternative indicators of achievement by the students of a particular state. Both of these tests can serve as measures of transfer of learning.

At this time, we know that the set of states without high-stakes tests perform better on the ACT and SAT. We do not know, however, how performance on the ACT and SAT tests changed after high school graduation exams were implemented in the 18 states that have introduced high-stakes testing policies. The objective of the first section of this inquiry is to answer this question.

There are, however, limitations to using these measures. For example, students who take the ACT and SAT are college-bound students and do not represent all students in a given state. But in 2001 38% and 45% of all graduating seniors took the ACT and the SAT tests, respectively. Although the sample of students is not representative, we can still use these scores to assess how high-stakes tests affected an average of approximately 2 out of every 5 students across the nation. Additionally, because participation rates vary by state we can use state participation rates to assess how in some states high-stakes tests affected the academic performance of more than 75% of graduating seniors.

It should be noted, as well, that some states are ACT states or states in which the majority of high school seniors take the ACT. Other states are SAT states or states in which the majority of high school seniors take the SAT. In Mississippi, for example, only 4% of high school seniors took the SAT in 2001 but in that same year 89% of high school seniors took the ACT. This would make Mississippi an ACT state. Whether states with high-stakes tests are ACT or SAT states should be taken into consideration to help us understand the sample of students who are taking the tests. If within Mississippi only 4% of high school seniors took the SAT it can be assumed that those students were probably among the brightest or most ambitious high school seniors in Mississippi. These students probably take the

SAT because they were seeking out-of-state universities. Conversely, if 89% of high school seniors took the ACT, it can be assumed that those students were probably a bit less talented or ambitious seniors, predominantly students trying to meet the requirements of the universities within the state of Mississippi. It is likely, however, that this sample also includes those seeking entrance to out of state universities that accept ACT scores. The participation rates for each test helps to decipher whether different samples of college bound students performed differently.

It should also be noted that the ACT and SAT tests are high-stakes tests. A student's score does influence to which colleges a student may apply and in which colleges a student may enroll. It seems likely, therefore, that students who take these tests are trying to achieve the highest scores possible. This would deflate arguments that students try harder on high school graduation exams than college entrance exams. If anything, the opposite might be true.

The purpose in the next two analyses is to assess how student learning changed in the domains represented by the ACT and SAT. Student scores and participation rates on these tests will be examined in each state after high-stakes high school graduation tests were implemented. Effects will be analyzed from the year in which the first graduating class was required to pass a high school graduation exam. It is also the purpose of the next two analyses to assess how high school seniors who are likely to be bound for out-of-state colleges, and seniors likely to be bound for in-state colleges, performed after high school graduation high-stakes exams were implemented.

American College Test (ACT)

The ACT data for each of the 18 states with high-stakes testing is included in Appendix B. Short-term, long-term, (Note 105) and overall achievement trends on the ACT were analyzed in the years following a states implementation of a high-stakes high school graduation exam. These analyses are summarized in Appendix B as well. The data and analysis for the state of Alabama, which we included as Figure 2, illustrated the way we examined each state's ACT data. A summary of those trends across the 18 states with the highest stakes is provided in Table 4.

Table 4
Results from the Analysis of ACT Scores (Note 106)

State	Effect after 1st HSGE		Effect after 2nd HSGE		Effect after 3rd HSGE		Effect after 4th HSGE		Overall Effects
	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	
Alabama	1984–'85 +0.1	1984–'92 +0.3	1992–'93 +0.1	1992–'01 – 0.1					Positive
Florida	n/a	1980–'89 – 0.4	1989–'90 – 0.1	1989–'95 – 0.2	1995–'96 – 0.3 (+2%)	1995–'01 – 0.6 (+5%)			Negative
Georgia	1983–'84 +0.2	1983–'94 – 0.5	1994–'95 – 0.1 (0%)	1994–'01 – 0.6 (0%)					Negative
Indiana	1999–'00 +0.2 (–1%)	1999–'01 +0.2 (–1%)							Positive
Louisiana	1990–'91 0	1990–'01 – 0.2							Negative
Maryland	1986–'87 +0.1	1986–'01 – 0.6							Negative
Minnesota	1999–'00 – 0.1 (0%)	1999–'01 0 (0%)							Negative
Mississippi	1988–'89 0	1988–'01 – 0.4							Negative
Nevada	1980–'81 – 0.1	1980–'84 +0.1	1984–'85 – 0.3	1984–'91 +0.1	1991–'92 +0.2	1991–'98 +0.1	1998–'99 +0.1 (–1%)	1998–'01 – 0.1 (–5%)	Positive
New Jersey	1983–'84 +0.3	1983–'86 – 1.4	1986–'87 – 0.2	1986–'94 – 0.1	1994–'95 – 0.5 (–1%)	1994–'01 – 0.5 (–1%)			Negative
New Mexico	1989–'90 +0.1	1989–'01 – 0.5							Negative
New York	1984–'85 – 0.2	1984–'94 – 0.5	1994–'95 +0.1 (–1%)	1994–'01 +0.4 (–6%)					Negative
North Carolina	1979–'80 n/a	1980–'97 – 1.1	1997–'98 +0.1 (0%)	1997–'01 +0.4 (0%)					Negative
Ohio	1993–'94 +0.1	1993–'01 +0.1							Positive
South Carolina	1989–'90 +0.1	1989–'01 – 0.5							Negative

249

Tennessee	1985–'86 +0.3	1985–'97 – 0.3	1997–'98 +0.1 (–7%)	1997–'01 +0.3 (–6%)		Positive
Texas	1986–'87 +0.2	1986–'91 +0.7	1991–1'92 0	1991–'01 0		Positive
Virginia	1985–'86 – 0.1	1985–'01 – 1.3				Negative

From Table 4, looking at all the states simultaneously, and in comparison to the nation, we can evaluate short-term, long-term, and the overall effects of high stakes testing policies.

Short-term effects. In the short term, ACT gains were posted 1.6 times more often than losses after high school graduation exams were implemented. Short-term gains were evident sixteen times, losses were evident ten times, and no apparent effects were evident three times. But the gains and losses that occurred were partly artificial, because the states' short-term changes in scores were correlated ($-0.51 < r < 0.13$) (Note 107) to the states short-term changes in participation rates. This modest negative correlation informs us that if the participation rate in ACT testing went down then the scores on the ACT went up, and vice versa. Under these circumstances it is hard to defend the thesis that there are reliable short-term gains from high-stakes tests.

Long-term effects. In the long term, and also in comparison to the nation, ACT losses were posted 1.9 times more often than gains after high school graduation exams were implemented. Long-term gains were evident ten times, losses were evident nineteen times, and no apparent effects were evident two times. These gains and losses were "real" given that the states' long-term changes in score were unrelated ($r = -0.18$) (Note 108) to the states' long-term changes in participation rates.

Overall effects. In comparison to the rest of the nation, negative ACT effects were displayed 2 times more often than positive effects after high-stakes high school graduation exams were implemented. Six states displayed overall positive effects, while twelve states displayed overall negative effects. In this data set overall losses or gains were unrelated to whether the percentage of students participating in the ACT increased or decreased.

Assuming that the ACT can serve as an alternative measure of the same or a similar domain as a state's high-stakes achievement tests, there is scant evidence of learning. Although states may demonstrate increases in scores on their own high-stakes tests, it appears that transfer of learning is not a typical outcome of their high-stakes testing policy. Sixty-seven percent of the states that use high school graduation exams posted *decreases* in ACT performance after high school graduation exams were implemented. These decreases were unrelated to whether participation rates increased or decreased at the same time. On average, the college-bound students in states with high school graduation exams decreased in levels of academic achievement as measured by the ACT.

One additional point about the ACT data needs to be made. In ACT states (states in which more than 50% of high school seniors took the ACT) students who are thought to be headed for in-state colleges were just slightly (1.3 times) more likely to post negative effects on the ACT. In SAT states (states in which less than 50% of high school seniors took the ACT) the students who are more likely bound for out-of-state colleges were 2.7 times more likely to post negative effects on the ACT. If anything, high school graduation exams hindered the performance of the brightest and most ambitious of the students bound for out-of-state colleges. Seventy-three percent of the states in which less than 50% of students take the ACT posted overall losses on the ACT.

Analysis of ACT Participation Rates. (Note 109) Just as ACT scores were used as indicators of academic achievement, ACT participation rates were used as indicators of the rates by which students in each state were planning to go to college. Arguably, if high school graduation exams increased academic achievement in some broad and general sense, an increase in the number of students pursuing a college degree would be noticed. An indicator of that trend would be increased ACT participation rates over time. So we examined changes in the rates by which students participated in ACT testing after the year in which the first graduating class was required to pass a high school graduation exam and for which data were available. These results are presented in Table 5.

Table 5
Results from the Analysis of ACT Participation Rates

State	Year in which students had to pass 1st HSGE to graduate	Change in % of students taking the ACT 1994–2001 as compared to the nation*	Overall Effects
Alabama	1985	+9%	Positive
Florida	1979	+4%	Positive
Georgia	1984	0%	Neutral
Indiana	2000	–1%	Negative
Louisiana	1991	+5%	Positive

Maryland	1987	–1%	Negative
Minnesota	2000	0%	Neutral
Mississippi	1989	+14%	Positive
Nevada	1981	–6%	Negative
New Jersey	1985	–1%	Negative
New Mexico	1990	0%	Neutral
New York	1985	–6%	Negative
North Carolina	1980	+2%	Positive
Ohio	1994	+2%	Positive
South Carolina	1990	+15%	Positive
Tennessee	1986	+10%	Positive
Texas	1987	–2%	Negative
Virginia	1986	+4%	Positive

*1999–2001 data were used for Indiana and Minnesota.

From this analysis we learn that from 1994–2001 ACT participation rates, as compared to the nation, increased in 50% of the states with high school graduation exams. When compared to the nation, participation rates increased in nine states, decreased in six states, and stayed the same in three states. Thus there is scant support for the belief that high-stakes testing policies within a state have an impact on the rate of college attendance.

The Scholastic Achievement Test (SAT)

The SAT data for each of the 18 states with high-stakes testing is included in Appendix C. Short-term, long-term, and overall achievement trends were analyzed following the states' implementation of their high-stakes high school graduation exam and these analyses are summarized in Appendix C, as well. The state of Florida was randomly chosen from this data set to illustrate what a time series for the SAT looks like. These data are provided in Figure 3. A summary of those trends across the 18 high-stakes testing states is provided in Table 6.
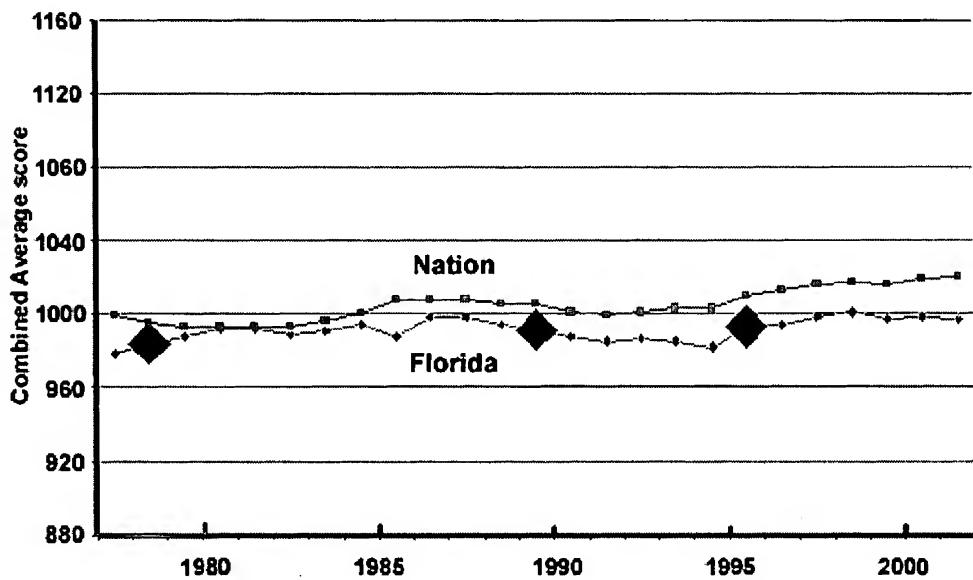

Figure 3. Florida: SAT scores

Florida implemented its 1st high school graduation exam in 1976. It was a prerequisite for graduation that first affected the class of 1979. Florida's 2nd exam first affected the class of 1990 and its 3rd exam the class of 1996 – see points of intervention (diamonds) enlarged to signify the year *before* the 1st graduating class was required to pass each exam:

- From 1978–1979 Florida gained 6 points on the nation.
- From 1978–1989 Florida lost 4 points to the nation.
- From 1989–1990 Florida gained 2 points on the nation.
- From 1989–1995 Florida lost 2 points to the nation.
- From 1995–1996 Florida lost 2 points to the nation.
- From 1995–2001 Florida lost 6 points to the nation.

Table 6
Results from the Analysis of SAT Scores Across the States (Note 110)

State	Effect after 1st HSGE		Effect after 2nd HSGE		Effect after 3rd HSGE		Effect after 4th HSGE		Overall Effects
	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term	
Alabama	1984–85 +13	1984–92 +23	1992–93 +7 (0%)	1992–01 +4 (–2%)					Positive
Florida	1978–79 +6	1978–89 –4	1989–90 +2	1989–95 –2	1995–96 –2 (0%)	1995–01 –6 (+2%)			Negative
Georgia	1983–84 0	1983–94 +21	1994–95 –2 (+1%)	1994–2001 +10 (–5%)					Positive
Indiana	1999–00 +2	1999–01 +2							Positive
Louisiana	1990–91 +3	1990–01 +19							Positive
Maryland	1986–87 +3	1986–01 –6							Negative
Minnesota	1999–00 –12	1999–01 –19							Negative
Mississippi	1988–89 –13	1988–01 +7							Negative
Nevada	1980–81 +3	1980–84 –6	1984–85 –16	1984–91 –10	1991–92 +1 (+2%)	1991–98 –15	1998–99 +7	1998–01 –2	Negative
New Jersey	1983–84 –2	1983–86 +2	1986–87 +4	1986–94 +8	1994–95 –2 (0%)	1994–01 +1 (+7%)			Positive
New Mexico	1989–90 –3	1989–01 –29							Negative
New York	1984–85 –3	1984–94 –11	1994–95 –3 (–1%)	1994–2001 –6 (–2%)					Negative
North Carolina	1980–81 +7	1980–97 +32	1997–98 +3	1997–01 +10					Positive
Ohio	1993–94 +7	1993–01 –1							Positive
South Carolina	1989–90 +2	1989–01 +15							Positive
Tennessee	1985–86 –3	1985–97 +8	1997–98 +1	1997–01 –9 (–3%)					Negative
Texas	1986–87 –2	1986–91 +7	1991–92 –1 (0%)	1991–01 –8 (+6%)					Negative
Virginia	1985–86 +1	1985–01 –9							Negative

Short–term effects. Looking across all the states simultaneously, and in comparison to the nation, we see that in the short term, SAT gains were posted 1.3 times more often than losses after high school graduation exams were implemented. Short-term gains were posted seventeen times, losses were posted thirteen times, and no apparent effects were posted once. But the gains and losses that occurred were partly artificial because the states' short-term changes in scores were related ($-0.60 < r < 0.38$) to the states short-term changes in participation rates. The negative correlations inform us that if the participation rate in SAT testing went down the scores on the SAT went up, and vice versa. The modest positive correlations inform us that in a few cases if the participation rate in SAT testing went down the scores on the SAT went down, and vice versa. Under these circumstances it is hard to defend the thesis that there are any reliable short-term gains on measures of general learning associated with high-stakes tests.

Long-term effects. In the long term, and also in comparison to the nation, SAT losses were posted 1.1 times more often than gains after high school graduation exams were implemented. Long-term gains were evident fifteen times, and losses were evident sixteen times. These gains and losses were partly artificial, however, given that the states' long-term changes in score were negatively correlated ($r = -0.41$) to the changes in participation rates for taking the

BEST COPY AVAILABLE

SAT. The fewer students taking the test, the higher the SAT scores, and vice versa.

Overall effects. In comparison to the rest of the nation, negative SAT effects were posted 1.3 times more often than positive effects after high school graduation exams were implemented. Eight states displayed overall positive effects, while ten states displayed overall negative effects. But the gains or losses in score were related to increases and decreases in the percentage of students participating in the SAT. Thus it is hard to attribute any effects on the SAT to the implementation of high-stakes testing.

If we assume that the SAT is an alternative measure of the same or a similar domain as a state's own high-stakes achievement tests, then there is scant evidence of learning. Although states may demonstrate increases in scores on their own high-stakes tests, it appears that transfer of learning is not a typical outcome of their high-stakes testing policy. Fifty-six percent of the states that use high school graduation exams posted decreases in SAT performance after high school graduation exams were implemented. However, these decreases were slightly related to whether SAT participation rates increased or decreased at the same time. Thus, there is no reliable evidence that high-stakes high school graduation exams improve the performance of students who take the SAT. Gains and losses in SAT scores are more related to who participates in the SAT than the implementation of high school graduation exams.

One additional point about the SAT data needs to be made. In SAT states (states in which more than 50% of high school seniors took the SAT) students who are thought to be headed for in-state colleges were equally likely to post negative and positive effects on the SAT. In ACT states (states in which less than 50% of high school seniors took the SAT) the students who are more likely bound for out-of-state colleges were 1.7 times more likely to post negative effects on the SAT. If anything, high school graduation exams hindered the performance of the brightest and most ambitious of the students bound for out-of-state colleges. Sixty-three percent of the states in which less than 50% of students take the SAT posted overall losses on the SAT.

Analysis of SAT Participation Rates. Just as SAT scores were used as indicators of academic achievement, SAT participation rates were used as indicators of the rates by which students in each state were planning to go to college. Arguably, if high school graduation exams increased academic achievement in some broad and general sense, an increase in the number of students pursuing a college degree would be noticed. An indicator of that trend would be increased SAT participation rates. So we examined changes in the rates by which students participated in SAT testing after the year in which the first graduating class was required to pass a high school graduation exam and for which data were available. These results are presented in Table 7.

Table 7
Results from the Analysis of SAT Participation Rates

State	Year students must pass 1st HSGE to graduate	Change in % of students taking the SAT 1991–2001 as compared to the nation*	Overall Effects
Alabama	1985	–2%	Negative
Florida	1979	+3%	Positive
Georgia	1984	–2%	Negative
Indiana	2000	–1%	Negative
Louisiana	1991	–5%	Negative
Maryland	1987	–2%	Negative
Minnesota	2000	–1%	Negative
Mississippi	1989	–3%	Negative
Nevada	1981	+5%	Positive
New Jersey	1985	+4%	Positive
New Mexico	1990	–2%	Negative
New York	1985	–1%	Negative
North Carolina	1980	+5%	Positive
Ohio	1994	+2%	Positive
South Carolina	1990	–4%	Negative
Tennessee	1986	–2%	Negative
Texas	1987	+6%	Positive
Virginia	1986	+5%	Positive

* 1993–2001 data were used for Ohio and 2000–2001 data were used for Indiana and Minnesota. Participation rates were not available for 1998 and 1999.

From this analysis we learn that from 1991–2001 (1993–2001 in Ohio, and 2000–2001 in Indiana and Minnesota) SAT participation rates, as compared to the nation, fell in 61% of the states with high school graduation exams. Participation rates in the SAT increased in seven states and decreased in eleven states. There is scant support for the belief that high-stakes testing policies will increase the rate of college attendance. Students did not participate in the SAT testing program at greater rates after high-stakes high school graduation exams were implemented.

National Assessment of Educational Progress (NAEP)

Some may argue that using ACT and SAT scores to assess the effects of high school graduation exams is illogical because high school graduation exams are specifically intended to raise the achievement levels of those students who are the most likely to fail – the poor, in general, and poor racial minorities, in particular. These students do not take the ACT or SAT in great numbers. But the effects of high-stakes policies on these particular populations can be assessed with data from the National Assessment of Educational Progress.(Note 111)

The National Assessment of Educational Progress (NAEP), commonly known as 'the nation's report card," is the test administered by the federal government to monitor the condition of education in the nation's schools. NAEP began in 1969 as a national assessment of three different age or grade levels, for which students were randomly sampled and tested to provide information about the outcomes of the nation's various educational systems. In 1990 NAEP was expanded to provide information at the state level, allowing for the first time state-to-state comparisons.

States that volunteered to participate in NAEP could gauge how they performed in math and reading in comparison to each other and to the nation, overall. This way states could assess the effects of the particular educational policies they had implemented. Under President Bush's national education policy, however, states are required to take the NAEP because it is believed to be the most robust and stable instrument the nation has to gauge learning and educational progress across all states.(Note 112) The federal government believes, as we do, that NAEP exams can be used to asses transfer, that NAEP is an alternate measure of the domains that are assessed by each of the states.

Weaknesses of the NAEP. It is proper to acknowledge that the NAEP has a number of weaknesses influencing interpretations of the data that we offer below. First, state level NAEP data pertain only to 4th and 8th grade achievement. The national student data set includes 12th grade data as well, and some additional subjects are tested, but at the state levels, only 4th and 8th grade achievement is measured. Given these circumstances it is not logical to attempt an assessment of the effects of implementing a high school graduation exam, or any other exam that is usually administered at the high school level, by analyzing NAEP tests given at the 4th or 8th grade. On the other hand, it is not illogical to make the assumption that other state reform policies went into effect at or around the same time as high-stakes high school graduation exams were put into place, *including the use of other high-stakes tests at lower grade levels.*(Note 113) The usefulness of the NAEP analyses that follow rest on the assumption that states' other K–12 high-stakes testing policies were implemented at or around the same time as each state's high school graduation exam. Table 1 describes these policies, and these policies are elaborated on in Appendix A. Other researchers who have used NAEP data to draw conclusions about the effects of high-stakes tests have used this logic and methodology as well.(Note 114)

Secondly, the NAEP does not have stakes attached to it. Students who are randomly selected to participate do not have to perform their best. However, because each student only takes small sections of the test, students appear to be motivated to do well and the scores appear to be trustworthy.(Note 115)

Third, states like North Carolina have aligned their state-administered exams with the NAEP, making for state-mandated tests that are very similar to the NAEP.(Note 116) In such cases gains in score on the NAEP may be related to similarities in test content rather than actual increases in school learning. States that align their tests with the NAEP have an unfair advantage over other states that aligned their tests with their state standards, but such imitative forms of testing occur. State tests that look much like the NAEP will probably become more common now that President Bush is attempting to attach stakes to the NAEP, and this will, of course, make the NAEP much less useful as a yardstick to assess if genuine learning of the domains of interest is taking place.

Finally, when analyzing NAEP data it is important to pay attention to who is actually tested. The NAEP sampling plan uses a multi-stage random sampling technique. In each participating state, school districts are randomly sampled. Then, schools within districts are randomly sampled. And then, students within schools are randomly sampled. Once the final list of participants is drawn, school personnel sift through the list and remove students who they have classified as Limited English Proficient (LEP) or who have Individualized Education Plans (IEPs) as part of their special education programs. Local personnel are required to follow "carefully defined criteria" in making determinations as to whether potential participants are "capable of participating."(Note 117) In short, although the NAEP uses random sampling techniques, not all students sampled are actually tested. The exclusion of these students biases NAEP results.

Illusion from exclusion. Walter Haney found that exclusion rates explained gains in NAEP scores and vice versa. Texas, for example, was one state in which large gains in NAEP scores were heralded as proof that high-stakes tests do, indeed, improve student achievement. But Haney found that the percentages of students excluded from participating in the NAEP increased at the same time that large gains in scores were noted. Exclusion rates increased at both grade levels escalating from 8% to 11% at grade 4 and from 7% to 8% at grade 8 from 1992–1996.

254

Meanwhile, in contrast, exclusion rates declined at both grade levels at the national level during this same time period, decreasing from 8% to 6% at grade 4 and from 7% to 5% at grade 8. Haney, therefore, termed the score gains in Texas an "illusion arising from exclusion."(Note 118)

Unfortunately, however, such illusions from exclusions hold true across the other states that use high-stakes tests. For example, North Carolina was the other state in which large gains in NAEP scores were heralded as proof that high-stakes testing programs improve student achievement. On the 4th grade NAEP math test North Carolina recorded an average composite score of 212 in 1992 and an average composite score of 232 in 2000. The nation's composite score increased from 218 to 226 over the same time period. North Carolina gained 20 points while the nation gained 8, making for what would seem to be a remarkable 12-point gain over the nation, the largest gain made by any state. But North Carolina excluded 4% of its LEP and IEP students in 1992 and 13% of its LEP and IEP students in 2000. Meanwhile, the nation's exclusion rate decreased from 8% to 7% over the same time period. North Carolina excluded 9% more of its LEP and IEP students while the nation excluded 1% less making for a 10% divergence between North Carolina's and the nation's exclusion rates from 1992–2000. North Carolina's grade 4 math 1992–2000 exclusion rates increased 325% while the nation's exclusion rate decreased. In addition, North Carolina's grade 8 math 1992–2000 exclusion rates increased 467% while the nation's exclusion rate stayed the same.

There is little doubt that the relative gains posted by North Carolina were partly, if not entirely, artificial given the enormous relative increase in the rates by which North Carolina excluded students from participating in the NAEP. The Heisenberg Uncertainty Principle appears to be at work in both Texas and North Carolina, leading to distortions and corruptions of the data, giving rise to uncertainty about the meaning of the scores on the NAEP tests.

North Carolina and Texas, however, are not the only states in which exclusionary trends were observed. In states with high-stakes tests, between 0%–49% of the gains in NAEP scores can be explained by increases in rates of exclusion. Similarly, 0%–49% of the losses in score can be explained by decreases in rates of exclusion over the same years. (Note 119) The more recent the data, the more the variance in NAEP scores can be explained by changes in exclusion rates. In short, states that are posting gains are increasingly excluding students from the assessment. This is happening with greater frequency as time passes from one NAEP test to the next. That is, as the stakes attached to the NAEP become higher, the Heisenberg Uncertainty Principle in assessment apparently is having its effects, with distortions and corruptions of the assessment system becoming more evident.

The state scores on the NAEP math and reading tests, at grades 4 and 8, will be used in our analysis to test the effects on learning from using high-stakes tests in states that have implemented high-stakes high school graduation exams. Given that exclusion rates affect gains and losses in score, however, state exclusion rates will be presented along side the relative gains or losses posted by each state. In this way readers can make their own judgments about whether year-to-year gains in score are likely to be "true" or "artificial." The gains and losses in scores and exclusion rates have all been calculated in comparison to the pooled national data.

Analysis of NAEP Grade 4 Math Scores

For each state, after high-stakes tests were implemented, an analysis of NAEP mathematics achievement scores was conducted. The state of Georgia was randomly chosen to serve as an example of the analysis we did on the grade 4 NAEP math tests (see Figure 4). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs and the reform efforts that accompany them should affect learning in the different mathematics domains that make up the K–4 curriculum. NAEP is a test derived from the K–4 mathematics domains.
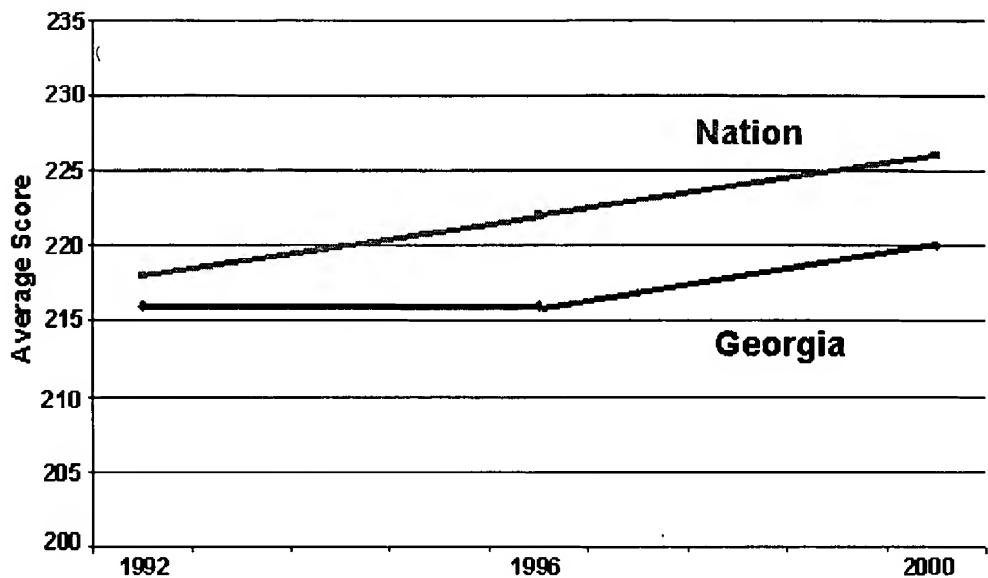
Figure 4. NAEP Math, Grade 4: Georgia

Trend lines and analytic comments for all the other states are included in Appendix D. A summary of these data across all 18 states is presented as Table 8.

Georgia implemented its 1st high school graduation exam in 1984. Assuming that other stakes attached to Georgia's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter:

- From 1992–1996 Georgia lost 4 points to the nation.
- From 1996–2000 Georgia gained 4 points, as did the nation.
- From 1992–2000 Georgia lost 4 points to the nation.

Table 8
Results from the Analysis of NAEP Math Grade 4 Scores

State	Year in which students had to pass 1st HSGE to graduate	1992–1996 Change in score	1992–1996 Change in % excluded	1996–2000 Change in score	1996–2000 Change in % excluded	1992–2000 Change in score	1992–2000 Change in % excluded	Overall Effects
Alabama	1985	0	+3%	+2	–1%	+2	+2%	Positive
Florida	1979	–2	n/a	n/a	n/a	n/a	n/a	Negative
Georgia	1984	–4	+4%	0	–1%	–4	+3%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	+1	+6%	+5	–1%	+6	+5%	Positive
Maryland	1987	–1	+6%	–2	0%	–3	+6%	Negative
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	+2	+3%	–1	–3%	+1	0%	Positive
Nevada	1981	n/a	n/a	–1	0%	n/a	n/a	Negative
New Jersey	1984	–4	n/a	n/a	n/a	n/a	n/a	Negative
New Mexico	1990	–3	+7%	–4	–1%	–7	+6%	Negative
New York	1985	0	+5%	0	+3%	0	+8%	Neutral
North Carolina	1980	+8	+5%	+4	+5%	+12	+10%	Positive
Ohio	1994	+2	n/a	+2	n/a	+4	+5%	Positive
South Carolina	1990	–3	+3%	+3	0%	0	+3%	Neutral
Tennessee	1986	+4	+4%	–3	–3%	+1	+1%	Positive
Texas	1987	+7	+4%	0	+4%	+7	+8%	Positive

| Virginia | 1986 | -2 | +4% | +4 | +3% | +2 | +7% | Positive |

The time period 1992–1996. From Table 8, in comparison to the nation as a whole, we see that the states that implemented high-stakes tests 1 or more years before 1996 posted losses 1.2 times more often than gains on the 1992–1996 grade 4 NAEP math tests. Six states posted gains, seven states posted losses, and two states posted no changes, as compared to the nation. Thus, only 40% of the states with high-stakes tests posted gains from 1992–1996. These gains and losses may be considered "real" given that the states' 1992–1996 changes in score were unrelated (r = 0) to the states' 1992–1996 exclusion rates.

The time period 1996–2000. Table 8 also reveals that on the 1996–2000 grade 4 NAEP math tests the states that implemented high-stakes tests 1 or more years before 2000 posted gains 1.2 times more often than losses, as compared to the nation. Six states posted gains, five states posted losses, and three states posted no changes as compared to the nation. Thus, only 43% of the states with high-stakes tests posted gains from 1996–2000. These gains and losses, however, were partly artificial since the states' 1996–2000 changes in score were positively correlated (r = 0.45) with the states' 1996–2000 exclusion rates.

The time period 1992–2000. Table 8 also reveals that states that implemented high-stakes tests 1 or more years before 2000 posted gains 2.7 times more often than losses. Another way to look at these data is to note that these states were 1.6 times more likely to show gains rather than losses or no changes on the grade 4 NAEP math tests over the time period from 1992–2000. Eight states posted gains, three states posted losses, and two states posted no changes as compared to the nation. Thus, gains were posted by 62% of the states with high-stakes tests from 1992–2000. But these gains and losses were partly artificial given that the states' 1992–2000 changes in score were positively correlated (r = 0.39) to the states' 1992–2000 exclusion rates. The higher the percent of students excluded, the higher the NAEP scores obtained by a state. Because of the correlation we found between exclusion rates and scores on the NAEP, there is uncertainty about the meaning of those improved scores.

The overall data set. In the years for which data were available, across all time periods, the implementation of high-stakes tests resulted in positive effects 1.3 times more often than negative effects on the grade 4 NAEP tests in mathematics. Eight states displayed positive effects, six states displayed negative effects, and two states displayed neutral effects. Thus, in comparison to national trends, 50% of the states with high-stakes tests posted positive effects but these gains and losses were partly artificial, given that the overall positive or negative changes in score were related to changes in the overall state exclusion rates.

In short, when compared to the nation as a whole, high-stakes testing policies did not usually lead to improvement in the performance of students on the grade 4 NAEP math tests between 1992 and 2000. Gains and losses were more likely to be related to who was excluded from the NAEP than to the effects of high-stakes testing programs in a state. In the 1992–1996 time period, when participation rates were unrelated to gains and losses, the academic achievement of students may have even been thwarted in those states where high-stakes testing was implemented. High-stakes tests within states probably had a differential impact on students from racial minority and economically disadvantaged backgrounds.

Analysis of NAEP Grade 8 Math Scores

For each state, after high-stakes tests had been implemented, an analysis of NAEP mathematics achievement scores was conducted. The state of Mississippi was randomly chosen to serve as an example of the analysis we did on the grade 8 NAEP math tests (see Figure 5). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different mathematics domains that make up the K–8 curriculum. NAEP is a test derived from the domains that make up the K–8 curriculum.
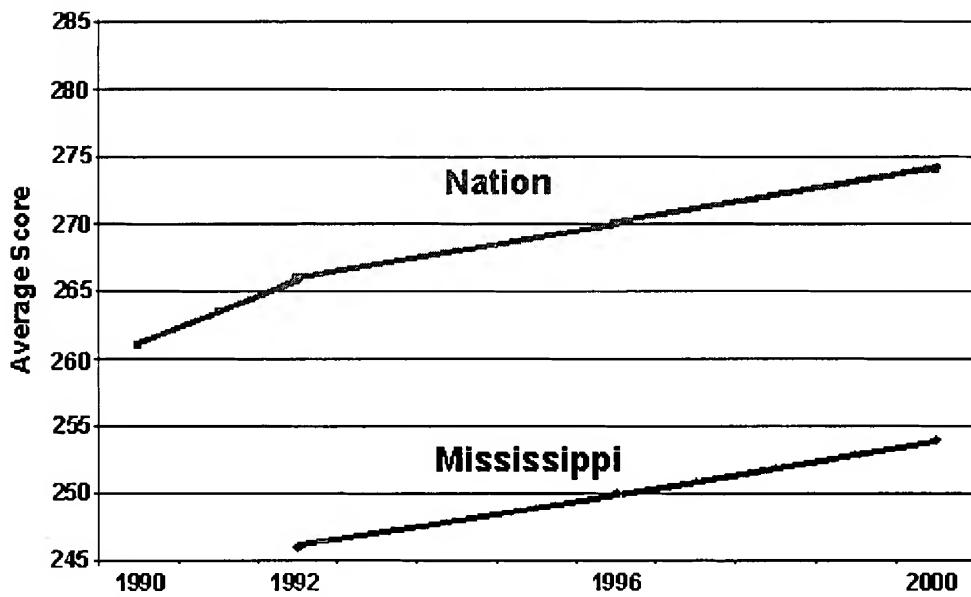
Figure 5. Mississippi – NAEP math grade 8

Mississippi implemented its 1st high school graduation exam in 1988. Assuming that the stakes attached to Mississippi's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter. From 1990–1992 Mississippi data were not available. From 1992–1996 Mississippi gained 4 points, as did the nation. From 1996–2000 Mississippi gained 4 points, as did the nation. From 1990–2000 Mississippi NAEP data were not available.

All other states' trend lines and analytic comments are included in Appendix E. A summary of these data across all 18 states is presented as Table 9.

Table 9
Results from the Analysis of NAEP Math Grade 8 Scores

State	Year in which students had to pass 1st HSGE to graduate	1990–92 Change in score	1990–1 92 Change in % excluded	1992–96 Change in score	1992– 96 Change in % excluded	1996–00 Change in score	1996– 00 Change in % excluded	1990–00 Change in score	1990–00 Change in % excluded	Overall Effects
Alabama	1985	–6	–2%	0	+4%	+2	–4%	–4	–2%	Negative
Florida	1979	–1	n/a	0	n/a	n/a	n/a	n/a	n/a	Negative
Georgia	1984	–5	0%	–1	+4%	0	–2%	–6	+2%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	–1	–2%	–2	+4%	+3	–2%	0	0	Neutral
Maryland	1987	–1	–2%	+1	+4%	+2	+2%	+2	+4%	Positive
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	0	n/a	0	+2%	n/a	n/a	n/a‾	n/a	Neutral
Nevada	1981	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Mexico	1990	–2	–3%	–2	+5%	–6	+2%	–10	+4%	Negative
New York	1985	0	0%	0	+2%	+2	+3%	+2	+5%	Positive
North Carolina	1980	+3	–2%	+6	+3%	+8	+8%	+17	+9%	Positive
Ohio	1994	–1	–1%	+3.5	n/a	+3.5	n/a	+6	+2%	Positive
South Carolina	1990	n/a	n/a	–4	+2%	+2	–1%	n/a	n/a	Negative
Tennessee	1986	n/a	n/a	+1	+1%	–4	–1%	n/a	n/a	Negative
Texas	1987	+2	–1%	+1	+4%	+1	–1%	+4	+2%	Positive

| Virginia | 1986 | -2 | -2% | -2 | +4% | +3 | +1% | -1 | +3% | Negative |

The time period 1990–1992. Table 9 reveals that, in comparison to the nation as a whole, states that implemented high-stakes tests one or more years before 1992 posted losses 4 times more often than gains on the 1990–1992 grade 8 NAEP math tests. Compared to the nation, two states posted gains, eight states posted losses, and two states posted no change. Over this time period gains on the NAEP tests were posted by 17% of the states with high-stakes tests. These gains and losses were "real" given that the states' 1990–1992 changes in score were unrelated (r = 0) to the states' 1990–1992 exclusion rates.

The time period 1992–1996. Table 9 also reveals that states that implemented high-stakes tests 1 or more years before 1996 were as likely to post gains as losses on the 1992–1996 grade 8 NAEP math tests. Five states posted gains, five states posted losses, and four states posted no changes as compared to the nation. Thus, from 1992–1996 only 36% of the states with high-stakes tests posted gains. These gains and losses were "real" given that the states' 1992–1996 changes in score were unrelated (r = 0) to the states' 1992–1996 exclusion rates.

The time period 1996–2000. Looking at the grade 8 NAEP math tests over the 1996–2000 time period we see that states that implemented high-stakes tests 1 or more years before 2000 posted gains 4.5 times more often than losses. Nine states posted gains, two states posted losses, and one state posted no changes, as compared to the nation. Thus, in the time period from 1996–2000 gains were posted by 75% of the states with high-stakes tests, but those NAEP scores were related to whether exclusion rates increased or decreased over the same time period, raising some uncertainty about the authenticity of these gains. Gains and losses during this time period must be considered partly artificial given that the states' 1996–2000 changes in score were positively related (r = 0.35) to the states' 1996–2000 exclusion rates.

The time period 1990–2000. Looking over the long term, states that implemented high-stakes tests one or more years before 2000 posted gains 1.3 times more often than losses on the 1990–2000 grade 8 NAEP math tests. Five states posted gains, four states posted losses, and one state posted no changes as compared to the nation. These gains and losses were partly artificial, however, given that the states' 1996–2000 changes in score were substantially related (r = 0.53) to the states' 1990–2000 exclusion rates.

Overall, across the years for which data were available, the states that had implemented high-stakes tests displayed negative effects 1.4 times more often than positive effects. Five states displayed positive effects, seven states displayed negative effects, and two states displayed neutral effects. Another way of interpreting these data is that 36% of the states with high-stakes tests posted positive effects from 1990–2000 on the grade 8 NAEP math examinations, while losses were posted by 50% of the states with high-stakes tests over this same time period. These gains and losses were partly artificial, however, given that the overall positive or negative changes in score were related to overall exclusion rates.

In short, there is no compelling evidence that high-stakes testing policies have improved the performance of students on the grade 8 NAEP math tests. Gains were more related to who was excluded from the NAEP than to whether there were high-stakes tests being used or not. If anything, the weight of the evidence suggests that high-stakes tests thwarted the academic achievement of students in these states.

Analysis of the Grade 4 NAEP Reading Scores

For each state, after high-stakes tests had been implemented, an analysis of NAEP reading achievement scores was conducted. The state of Virginia was randomly chosen to serve as an example of the analysis we did on the grade 4 NAEP reading tests (see Figure 6). The logic of this analysis rests on two assumptions. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of reading that make up the K–4 curricula. NAEP is a test derived from the various domains that constitute the K–4 reading curriculum.
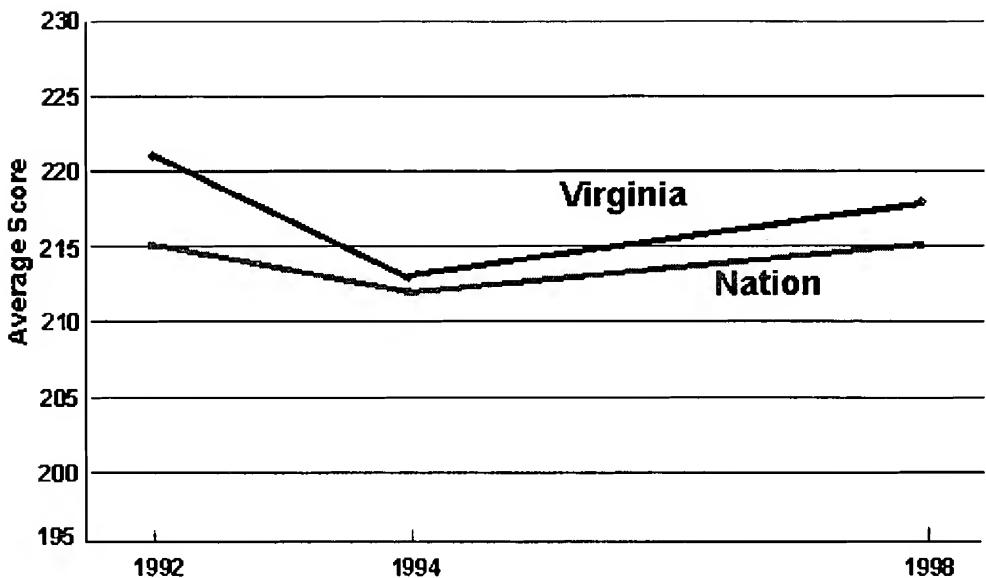
Figure 6. Virginia – NAEP reading grade 4

Virginia implemented its 1st high school graduation exam around 1981. Assuming that the stakes attached to Virginia's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter 1) From 1992–1994 Virginia lost 5 points to the nation; 2) From 1994–1998 Virginia gained 2 points on the nation; 3) From 1992–1998 Virginia lost 3 points to the nation. Trend lines and analytic comments for all other states are included in Appendix F. A summary of these data across all 18 states is presented as Table 10.

Table 10
Results from the analysis of NAEP reading grade 4 scores

State	Year in which students had to pass 1st HSGE to graduate	1992–94 Change in score	1992–94 Change in % excluded	1994–98 Change in score	1994–98 Change in % excluded	1992–98 Change in score	1992–98 Change in % excluded	Overall Effects
Alabama	1985	+4	–1%	0	+4%	+4	+3%	Positive
Florida	1979	0	+1%	–1	–1%	–1	0%	Negative
Georgia	1984	–2	0%	0	+2%	–2	+2%	Negative
Indiana	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Louisiana	1991	–4	+2%	+4	+7%	0	+9%	Neutral
Maryland	1987	+2	0%	+2	+3%	+4	+3%	Positive
Minnesota	2000	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mississippi	1989	+6	+1%	–1	–2%	+5	–1%	Positive
Nevada	1981	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a	n/a	n/a	n/a	n/a
New Mexico	1990	–3	0%	–2	+3%	–5	+3%	Negative
New York	1985	0	+2%	+1	0%	+1	+2%	Positive
North Carolina	1980	+5	+1%	0	+6%	+5	+7%	Positive
Ohio	1994	n/a	n/a	n/a	n/a	n/a	n/a	n/a
South Carolina	1990	–4	+1%	+4	+5%	0	+6%	Neutral
Tennessee	1986	+4	+1%	–4	–1%	0	0%	Neutral
Texas	1987	+2	+3%	+2	+3%	+4	+6%	Positive
Virginia	1986	–5	+1%	+2	+2%	–3	+3%	Negative

The time period 1992–1994. We note in Table 10 that on the grade 4 reading test, during the time period 1992–1994, states that implemented high-stakes tests 1 or more years before 1994 posted gains 1.2 times more often than losses,

260

in comparison to the nation. Compared to national trends six states posted gains, five states posted losses, and two states posted no changes at all. Thus, only 46% of the states with high-stakes tests posted gains from 1992–1994. These gains and losses were "real" given that the states' changes in score for the time period 1992–1994 were virtually unrelated (r = –0.10) to the states' exclusion rates.

The time period 1994–1998. Table 10 also reveals that those states implementing high-stakes tests 1 or more years before 1998 posted gains 1.5 times more often than losses when compared to national trends. Six states posted gains, four states posted losses, and three states posted no changes when compared to the national trends. Thus, only 46% of the states with high-stakes tests from 1994–1998 posted gains. These gains and losses were partly artificial, however, given that the states' 1994–1998 changes in score were strongly correlated (r = 0.63) to the states' 1994–1998 exclusion rates.

The time period 1992–1998. Table 10 also informs us that states implementing high-stakes tests 1 or more years before 1998 posted gains 1.5 times more often than losses in comparison to the national trends during the time period 1992–1998. Six states posted gains, four states posted losses, and three states posted no changes in comparison to national trends. Thus, only 46% of the states with high-stakes tests posted positive effects from 1992–1998 on the NAEP grade 4 reading test. The gains and losses may be considered "real" given that the states' 1992–1998 changes in score were virtually unrelated (r = 0.11) to the states' changes in 1992–1998 exclusion rates.

In short, in comparison to the national trends, high-stakes tests did not improve the learning of students as judged by their performance on the NAEP grade 4 reading test. This was clearest in the time periods from 1992–1994 and from 1992–1998. The learning effects over these years were unrelated to the rates by which students were excluded from the NAEP. We note, however, that in 1998 75% of the states with high-stakes tests had 1998 exclusion rates that were higher than the nation. Given the typical positive (and substantial) correlation between increased exclusion rates and increased NAEP scores, states' gains and losses in score need to be carefully evaluated. If anything, in comparison to national trends, the academic achievement of students in states with high-stakes testing policies seemed to be lower, particularly for students from minority backgrounds.

NAEP Cohort Analyses

Another way of investigating growth in achievement on measures other than states' high-stakes tests is to look at each state's cohort trends on the NAEP.(Note 122) The NAEP analyses preceding this section gauged the achievement trends of different samples of students over time, for example, 4th graders in one year compared to a different group of 4th graders a few years later. There is a slight weakness with this approach because we must compare students in one year with a different set of students a few years later. We are unable to control for differences between the different groups or cohorts of students.(Note 123) To compensate for this we did a cohort analysis, an analysis of the growth in achievement made by "similar" groups of students over time.

This is possible because NAEP uses random samples of students. Thus the 4th graders in 1996 should be representative of the same population of 8th graders tested four years later. Random sampling techniques made the groups of students similar enough so that the achievement effects made by the "same" (statistically the same) students can be tracked over time.(Note 124) Analyzing cohort trends in the 18 states with high-stakes tests helped assess the degree to which students increased in achievement as they progressed through school systems that were exerting more pressures for school improvement, including the use of high-stakes tests. We examined the growth of these students by tracking the relative changes in math achievement of 4th graders in 1996 to 8th graders in 2000, and by looking at the reading achievement of 4th graders in 1994 to that of 8th graders in 1998. The changes we record for each state are all relative to the national trends on the respective NAEP tests.

Cohort Analysis of NAEP Mathematics Scores: Grade 4 (1996) to Grade 8 (2000)

The state of New York was randomly chosen to serve as an example of the analysis we did for the NAEP mathematics cohort over the years 1996 to 2000 (see Figure 7). The logic of this analysis rests on the same two assumptions as previous NAEP analyses. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of mathematics from which NAEP is derived.
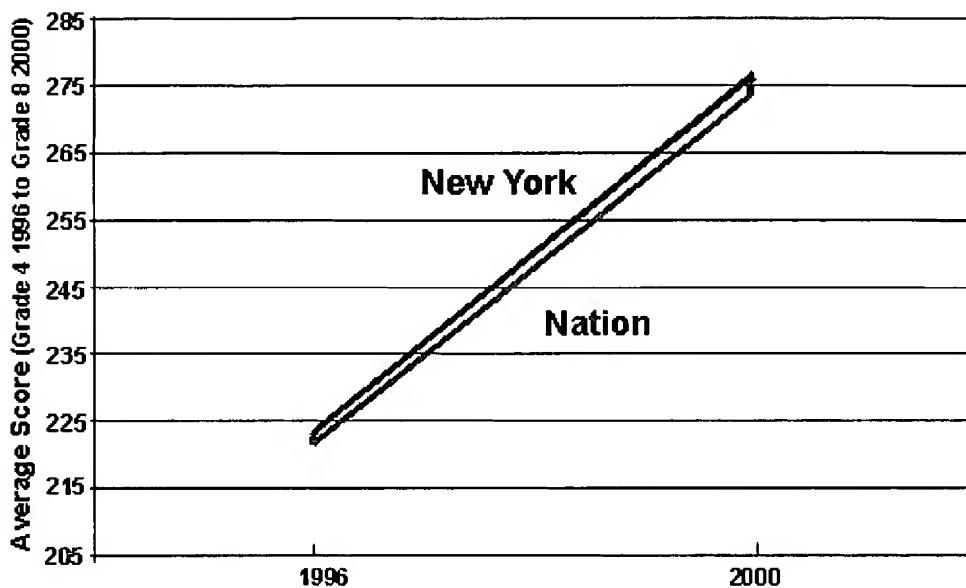
Figure 7. New York by cohort: NAEP math grade 4 1996 to grade 8 2000

New York implemented its 1st high school graduation exam around 1981. Assuming that the stakes attached to New York's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter, from 4th grade in 1996 to 8th grade in 2000 New York gained 1 point on the nation. Trend lines and analytic comments for all other states are included in Appendix G. A summary of these data across all 18 states is presented as Table 11.

Table 11
Results from the Analysis of NAEP Math Cohort Trends

State	Year in which students had to pass 1st HSGE to graduate	Change in score from grade 4 1996 to grade 8 2000	Change in % excluded from grade 4 1996 to grade 8 2000	Overall Effects 1996–00
Alabama	1985	–2	–2%	Negative
Florida	1979	n/a	n/a	n/a
Georgia	1984	–2	–1%	Negative
Indiana	2000	n/a	n/a	n/a
Louisiana	1991	–2	–3%	Negative
Maryland	1987	+4	+2%	Positive
Minnesota	2000	n/a	n/a	n/a
Mississippi	1989	–6	0%	Negative
Nevada	1981	–1	0%	Negative
New Jersey	1984	n/a	n/a	n/a
New Mexico	1990	–6	–1%	Negative
New York	1985	+1	+4%	Positive
North Carolina	1980	+4	+6%	Positive
Ohio	1994	n/a	n/a	n/a
South Carolina	1990	+1	0%	Positive
Tennessee	1986	–8	–2%	Negative
Texas	1987	–6	–1%	Negative
Virginia	1986	+3	+2%	Positive

The 1996–2000 cohort. From 1996 to 2000 cohorts of students moving from 4th to 8th grade in states that had implemented high-stakes tests in the years before 2000 posted losses 1.6 times more often than gains. In comparison to the national trends five states posted gains, and eight states posted losses. Said differently, in comparison to the nation, 62% of the states with high-stakes tests posted losses as their students moved from the 4th grade 1996 NAEP to the 8th grade 2000 NAEP. These gains and losses, however, were partly artificial because gains and losses in score for the cohorts in the various states were strongly correlated (r = 0.70) with overall exclusion rates. This cohort

262

analysis finds no evidence of gains in general learning as a result of high-stakes testing policies.

Cohort Analysis of NAEP Reading Scores: Grade 4 (1994) to Grade 8 (1998)

The state of Tennessee was randomly chosen to serve as an example of the analysis we did for NAEP reading cohort over the years 1994 to 1998 (see Figure 8). The logic of this analysis rests on the same two assumptions made in the previous NAEP analyses. First, that high-stakes tests and other reforms were implemented in all grades at or around the same time, or soon after high-stakes high school graduation exams were implemented. Second, that such high-stakes test programs should affect learning in the different domains of reading from which NAEP is derived.
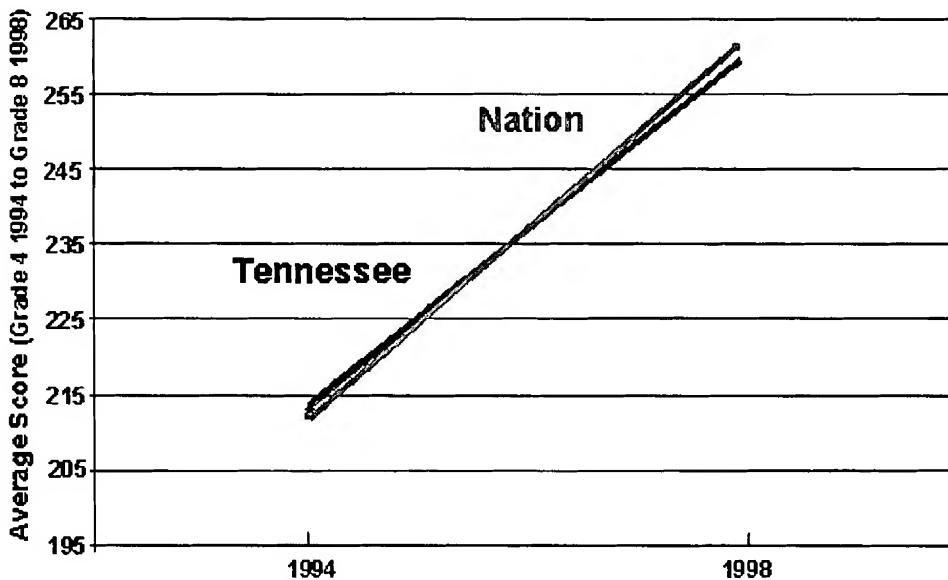

Figure 8. Tennessee by cohort: NAEP reading grade 4 1994 to grade 8 1998

Tennessee implemented its 1st high school graduation exam in 1982. Assuming that the stakes attached to Tennessee's Tennessee's K–8 tests (see Table 1) were attached at or around the same time or some time thereafter, from 4th grade in 1994 to 8th grade in 1998 Tennessee lost 3 points to the nation. Trend lines and analytic comments for all other states are included in Appendix H. A summary of these data across all 18 states is presented as Table 12.

Table 12
Results from the Analysis of NAEP Reading Cohort Trends

State	Year in which students had to pass 1st HSGE to graduate	Change in score from grade 4 1994 to grade 8 1998	Change in % excluded from grade 4 1994 to grade 8 1998	Overall Effects 1994–98
Alabama	1985	–2	+5%	Negative
Florida	1979	–1	–2%	Negative
Georgia	1984	+1	+4%	Positive
Indiana	2000	n/a	n/a	n/a
Louisiana	1991	+6	+6%	Positive
Maryland	1987	+3	+3%	Positive
Minnesota	2000	n/a	n/a	n/a
Mississippi	1989	–20	+4%	Negative
Nevada	1981	n/a	n/a	n/a
New Jersey	1984	n/a	n/a	n/a
New Mexico	1990	+4	+2%	Positive
New York	1985	+5	+4%	Positive
North Carolina	1980	+1	+7%	Positive
Ohio	1994	n/a	n/a	n/a
South	1990	+3	+2%	Positive

Carolina				
Tennessee	1986	–3	+1%	Negative
Texas	1987	+1	–1%	Positive
Virginia	1986	+4	+3%	Positive

The 1994–1998 cohort. In comparison to national trends, cohorts of students in states that implemented high-stakes tests in the years before 1998 posted gains 2.3 times more often than losses from the 4th to the 8th grade on the 1994 and the 1998 NAEP reading exams. Nine states posted gains, and four states posted losses. These gains and losses were "real" given that gains and losses in score were unrelated (r = 0) to overall exclusion rates.

Thus far in these analyses this is the only example we found of gains in achievement on a transfer measure that meet criteria of acceptability. As their students moved from the 4th grade in 1994 to the 8th grade in 1998, 69% of the states with high-stakes tests posted gains on the NAEP reading tests. Since these gains and losses were unrelated to increases and decreases in exclusion rates they appear to be "real" effects. To put these gains in context we note that in the states that showed increases in scores from 1994 to 1998, the average gain was 52 points. By any metric a 52-point gain is sizeable. But when these gains are compared to the national trends over the same time period, as shown in table12, we see that the gains in the states with high-stakes testing policies was, on average, only 3 points over the national trend. On the other hand, although fewer in number, the states that posted losses in comparison to the nation fell an average of 6.5 points. This figure is skewed, however, by the fact that Mississippi lost 20 points more than the nation did on the 4th to 8th grade reading NAEP from 1994-1998. In sum, these gains in the reading scores in states with high-stakes testing policies seem real but modest given the losses shown by other states with high-stakes testing policies.

Advanced Placement (AP) Data Analysis

The Advanced Placement (AP) program offers high school students opportunities to take college courses in a variety of subjects and receive credits before actually entering college. We used the AP data(Note 125) as another indicator of the effects of high-stakes high school graduation exams on the general learning and motivation of high school students. Using the AP exams as transfer measures and the AP participation rates as indicators of increased student preparation and motivation for college, we could inquire whether, in fact, high school graduation exams increased learning in the knowledge domains that are the intended targets of high-stakes testing programs.

The participation rates and rates by which students passed AP exams that are used in the following analyses were calculated by the College Board,(Note 126) administrators of the AP program. Gains or losses were assessed after the most recent year in which a new high school graduation exam was implemented or after 1995 – the first year for which these AP data were available.

Table 13 presents for each state the percentages of students who passed AP examinations with a grade of 3 or better after high school graduation exams were implemented. As we worked, however, it became apparent that fluctuations in participation rates were related (r = –0.30) to the percent of students passing AP exams with a grade 3 or better. If participation rates in a state decreased, the percent of students who passed AP exams usually increased and vice versa. To judge the effect of this interaction, and in comparison to the nation, the percent change in students who passed the AP examination is presented along with the percent change in students who participated in AP exams during the time period 1995-2000. If an increase in one corresponded to a decrease in the other, caution in making judgments about the effects is required.

North Carolina was randomly chosen from the states we examine to be the example for the AP analysis. That data is presented in Figure 9. Trend lines and analytic comments for all other states are included in Appendix I and summarized in Table 13.
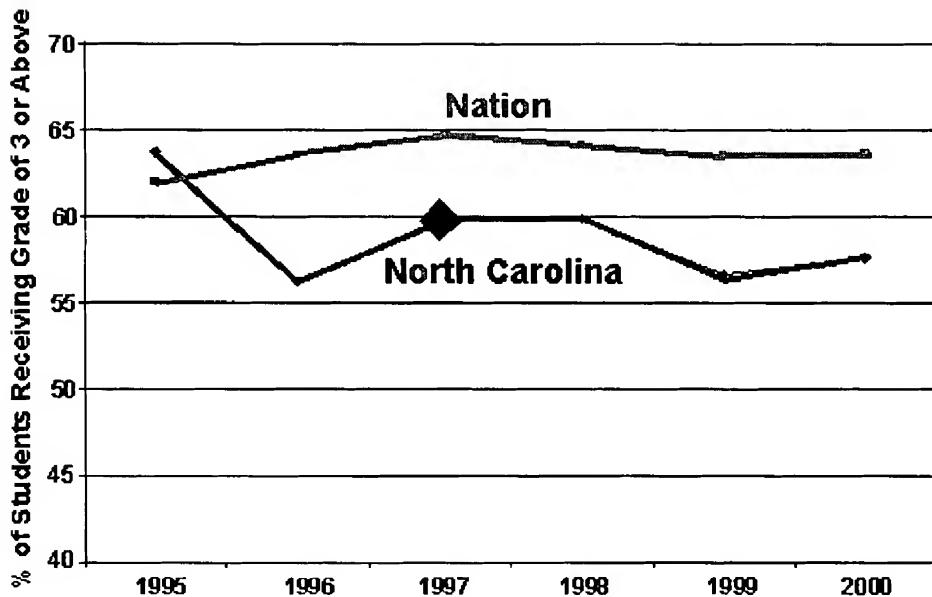
Figure 9. North Carolina: Percent passing AP examinations

North Carolina's 1st high school graduation exam first affected the class of 1980. North Carolina's second exam first affected the class of 1998. From 1995–2000 North Carolina lost 7.7 percentage points to the nation. From 1997–1998 North Carolina gained .5 percentage points on the nation. From 1997–2000 North Carolina lost 1.3 percentage points to the nation.

Table 13
Results from the Analysis of AP Scores and Participation Rates

State	Year in which students had to pass 1st HSGE to graduate	Change in % of students passing AP exams 1995–2000 as compared to the nation*	Change in % of students taking AP exams 1995–2000 as compared to the nation*	Overall Effects
Alabama	1985	+9.6%	-6.5%	Positive
Florida	1979	+3.9%	-0.5%	Positive
Georgia	1984	+6.8%	-1.4%	Positive
Indiana	2000	+1.9%	-0.4%	Positive
Louisiana	1991	+2.6%	-4.4%	Positive
Maryland	1987	+0.5%	+2.3%	Positive
Minnesota	2000	+0.6%	-1.6%	Positive
Mississippi	1989	-2.4%	-4.6%	Negative
Nevada	1981	+3.2%	-2.7%	Positive
New Jersey	1985	+1.7%	+2.0%	Positive
New Mexico	1990	-4.1%	-1.6%	Negative
New York	1985	+7.7%	+3.9%	Positive
North Carolina	1980	-7.7%	+0.9%	Negative
Ohio	1994	-3.3%	-2.6%	Negative
South Carolina	1990	-9.8%	-3.7%	Negative
Tennessee	1986	+5.8%	-1.8%	Positive
Texas	1987	-10.5%	+5.1%	Negative
Virginia	1986	-1.6%	+3.9%	Negative

*(Indiana and Minnesota, 1999–2000)

The time period 1995–2000. In comparison to national trends from 1995–2000, students in states with high school

graduation exams posted gains 1.6 times more than losses in the percentage of students passing AP exams with a score of 3 or better. Eleven states posted gains, and seven states posted losses. These gains and losses were partly artificial, however, given that gains and losses in the percentage of students passing AP exams were negatively correlated (r = –0.30) with the rate in which students participated in the AP program. The greater the percentage of students who participated in the AP program, the lower the percentage of students passing AP exams, and vice versa.

Compared to the national average participation rates fell in 67% of the states with high school graduation exams since 1995 (and since 1999 for Indiana and Minnesota). In comparison to the nation participation rates increased in six states and decreased in twelve states in the time period from 1995–2000.

Overall, 61% of the states with high-stakes tests posted gains in the rate by which students passed AP exams with a grade of 3 or better from 1995–2000 (1999–2000 in Indiana and Minnesota). But those increases and decreases in the percent passing AP exams were negatively correlated (r = –0.30) to whether participation rates increased or decreased at the same time. If we look at only those states where the participation rates did not seem to influence the percent passing AP exams(Note 127) as the overall correlation suggests it typically does, only Maryland (+), Mississippi (–), New Jersey (+), New Mexico (–), New York (+), Ohio (–), and South Carolina (–) posted "true" effects, 57% of which were negative.

The special case of Texas. Texas, as has been mentioned, received attention as one of two states in which high-stakes tests purportedly improve achievement. Dramatic gains in the rates of students enrolled in AP courses were among several state indicators of achievement provided by the state in support of their academic gains. But another educational policy was put into effect around the same time as the high-stakes testing program was implemented in that state.(Note 128) The Texas state legislature substantially reduced the cost of taking AP courses and the accompanying exams.(Note 129) This highly targeted policy may have helped increase enrollments in AP courses in Texas much more than their high-stakes testing program. So the substantial drop in the percent passing the test is difficult to assess since many more students took the AP tests. As we have seen, as a greater percentage of students in a state take the test the scores go down, and as a smaller percentage of students take the test scores go up. Inferences about the meaning of test scores become more uncertain when participation rates are not steady from one testing year to another.

In conclusion, when we use the national data on AP exams as a comparison for state AP data, and we use the percent of students passing the various AP exams as an indicator of learning in the domains of interest, we find no evidence of improvement associated with high-stakes high school graduation exams. When controlling for participation rates there even appeared to be a slight decrease in the percent of students who passed AP examinations. Further, in the states under study, high-stakes high school graduation exams did not result in an increase in the numbers of students preparing to go to college, as indicated by the percent of students who participated in AP programs from 1995–2000.

Conclusion

If we assume that the ACT, SAT, NAEP and AP tests are reasonable measures of the domains that a state's high-stakes testing program is intended to affect, then we have little evidence at the present time that such programs work. Although states may demonstrate increases in scores on their own high-stakes tests, transfer of learning is not a typical outcome of their high-stakes testing policy.

The ACT data. Sixty-seven percent of the states that use high school graduation exams posted *decreases* in ACT performance after high school graduation exams were implemented. These decreases were unrelated to whether participation rates increased or decreased at the same time. On average, as measured by the ACT, college-bound students in states with high school graduation exams decreased in levels of academic achievement. Moreover, participation rates in ACT testing, as compared to the nation, increased in nine states, decreased in six states, and stayed the same in three states. If participation rates in the ACT program serve as an indicator of motivation to attend college, then there is scant support for the belief that high-stakes testing policies within a state have such an impact.

The SAT data. Fifty six percent of the states that use high-stakes high school graduation exams posted decreases in SAT performance after those exams were implemented. However, these decreases were slightly related to whether SAT participation rates increased or decreased over the same time period. Thus, there is no reliable evidence of high-stakes high school graduation exams improving the performance of students who take the SAT. Gains and losses in SAT scores are more strongly correlated to who participates in the SAT than to the implementation of high school graduation exams. Moreover, SAT participation rates, as compared to the nation, fell in 61% of the states with high school graduation exams. If these participation rates serve as an indicator for testing the belief that high-stakes testing policies will prepare more students or motivate more students to attend college, then there is scant support for such beliefs. Students did not participate in the SAT testing program at greater rates after high-stakes high school graduation exams were implemented.

The NAEP mathematics data. High-stakes testing policies did not usually improve the performance of students on the grade 4 NAEP math tests. Gains and losses were more related to who was excluded from the NAEP than the effects of high-stakes testing programs in a state. However, during the 1992–1996 time period, when exclusion rates were unrelated to gains and losses in scores, mathematics achievement decreased for students in states where high-stakes

testing had been implemented. High-stakes testing policies did not consistently improve the performance of students on the grade 8 NAEP math tests. Gains were more strongly correlated to who was excluded from the NAEP than to whether or not high-stakes tests were used. If anything, the weight of the evidence suggests that students from states with high-stakes tests did not achieve as well on the grade 8 NAEP mathematics tests as students in other states.

The NAEP reading data. High-stakes testing policies did not consistently improve the general learning and competencies of students in reading as judged by their performance on the NAEP grade 4 reading test. This was clearest in the time periods from 1992–1994 and over the time span of from 1992–1998. The learning effects over these years were unrelated to the rates by which students were excluded from the NAEP. By 1998, however, 75% of the states with high-stakes tests had exclusion rates higher than the national average. These exclusionary policies were probably the reason for the apparent increases in achievement in several states. As the NAEP tests become more important in our national debates about school success and failure the effects of the Heisenberg Uncertainty Principle, as applied to the social sciences, seems to be evident. When these exclusion rates are taken into account, in comparison to national trends, the reading achievement of students in states with high-stakes testing policies appeared lower, particularly for students from minority backgrounds.

The NAEP cohort data. Sixty-two percent of the states with high-stakes tests posted losses on the NAEP mathematics exams as a cohort of their students moved from the 4th grade in 1996 to the 8th grade in the year 2000. These gains and losses, however, must be considered artificial to some extent because of the very strong relationship of overall exclusion rates to the gains and losses that were recorded. This cohort analysis finds no evidence of gains in general mathematics knowledge and skills as a result of high-stakes testing policies.

For the cohort of students moving from the 4th to the 8th grade and taking the 1994 and the 1998 NAEP reading exams, gains in scores were posted 2.3 times more often than losses in the states with high-stakes testing policies. Nine states (69%) posted gains, and four states (31%) posted losses. These gains and losses were "real" given that gains and losses in score were unrelated to overall NAEP exclusion rates. While not reflecting unequivocal support for high-stakes testing policies, this is the one case of gains in achievement on a transfer measure among the many analyses we did for this report. It is also true that over this time period many reading curriculum initiatives were being implemented throughout the country, as reading debates became heated and sparked controversy. Because of that it is not easy to attribute the gains made for the NAEP reading cohort to high-stakes testing policies. Our guess is that the reading initiatives and the high-stakes testing polices are entangled in ways that make it impossible to learn about their independent effects.

The AP data. High-stakes high school graduation exams do not improve achievement as indicated by the percent of students passing the various AP exams. When participation rates were controlled there was a decrease in the percent of students who passed AP examinations. Further, in the states with high-stakes high school graduation exams there was no increase in the numbers of students preparing to go to college, as indicated by the percent of students who chose to participate in AP programs from 1995–2000.

Final thoughts

. What shall we make of all this? At the present time, there is no compelling evidence from a set of states with high-stakes testing policies that those policies result in transfer to the broader domains of knowledge and skill for which high-stakes test scores *must* be indicators. Because of this, the high-stakes tests being used today do not, as a general rule, appear valid as indicators of genuine learning, of the types of learning that approach the American ideal of what an educated person knows and can do. Moreover, as predicted by the Heisenberg Uncertainty Principle, data from high-stakes testing programs too often appear distorted and corrupted.

Both the uncertainty associated with high-stakes testing data, and the questionable validity of high-stakes tests as indicators of the domains they are intended to reflect, suggest that this is a failed policy initiative. High-stakes testing policies are not now and may never be policies that will accomplish what they intend. Could the hundreds of millions of dollars and the billions of person hours spent in these programs be used more wisely? Furthermore, if failure in attaining the goals for which the policy was created results in disproportionate negative affects on the life chances of America's poor and minority students, as it appears to do, then a high-stakes testing policy is more than a benign error in political judgment. It is an error in policy that results in structural and institutional mechanisms that discriminate against all of America's poor and many of America's minority students. It is now time to debate high-stakes testing policies more thoroughly and seek to change them if they do not do what was intended and have some unintended negative consequences, as well.

Notes

1. Haladyna, Nolen, & Haas, 1991.

267

2. Figlio & Lucas, 2000.

3. Kreitzer, Madaus, & Haney, 1989.

4. Bracey, 1995; Heubert & Hauser, 1999; and Kreitzer, Madaus, & Haney, 1989.

5. Linn, 2000 and Serow, 1984.

6. U.S. Department of Education, 1983 and Bracey, 1995.

7. U.S. Department of Education, 1983.

8. Berliner & Biddle, 1995.

9. Quality Counts, 2001.

10. McNeil, 2000; Orfield & Kornhaber, 2001; Paris, 2000; Sacks, 1999; and Sheldon & Biddle, 1998.

11. Madaus & Clarke, 2001 and Campbell, 1975.

12. All of the following statistics come from extensive interviews conducted with knowledgeable testing personnel throughout the United States, and Quality Counts, 2001.

13. Administrative bonuses, 2001.

14. Neufeld, 2000.

15. Folmar, 2001.

16. Commission on Instructionally Supportive Testing, 2001.

17. California, Delaware, Michigan, Missouri, Nevada, and Ohio give scholarships to students for high performance on state mandated exams. See Quality Counts, 2001.

18. Durbin, 2001 and Ross, 2001.

19. Thanks to Professor J. Ryan, Arizona State University, for suggesting we investigate this story.

20. National Governor's Association, 2000; "Civil rights coalition," 2000; and "Using tobacco settlement revenues," 1999.

21. Heller, 1999. See also Durbin, 2001; Ross, 2001; and Swope & Miner, 2000.

22. "Civil rights coalition," 2000.

23. Heller, 1999.

24. Delaware, Ohio, South Carolina, and Texas have plans to promote students using test scores by the year 2003. Interview data and Quality Counts, 2001.

25. Florida implemented its first minimum competency test for the class of 1979, North Carolina implemented its first minimum competency test for the class of 1980, and Nevada implemented its first minimum competency test for the class of 1981.

26. U.S. Department of Education, 1983.

27. States that currently use high school graduation exams to grant or withhold diplomas are Alabama, Florida, Georgia, Indiana, Louisiana, Maryland, Minnesota, Mississippi, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, South Carolina, Tennessee, Texas, and Virginia. Hawaii used a test until 1999 and has plans to implement a different exam in 2007.

28. States that are developing high school exit exams are Alaska, Arizona, California, Delaware, Hawaii, Massachusetts, Utah, Washington, and Wisconsin.

29. Data illustrated in this chart were collected through telephone interviews and cross-checked with information

provided in Quality Counts, 2001. States are counted the year the first graduating class was (or will be) affected by the state's first high school high-stakes graduation exam. For example, since the class of 1987 was the first class that had to pass the TEAMS in Texas, Texas was defined as a state with a high school exit exam in 1987.

30. Percentages were calculated using 1997 National Center for Education Statistics finance data available: http://nccs.ed.gov/. . Data were adjusted for cost of living.

31. See Elazar's classification of state's governmental traditions of centralism and localism in Elazar, 1984. Hawaii and Alaska were not included in his analyses so were not included in these calculations.

32. These numbers were calculated using 2000 Census Bureau data available: http://www.ccnsus.gov/.

33. Ibid.

34. In the West, Nevada has a high school graduation exam and Alaska, California, Utah, and Washington have exams in progress (5/10 western states).

35. These numbers were calculated using 1999 Census Bureau data available: http://www.census.gov/.

36. New Mexico, Louisiana, California, Mississippi, New York, Alabama, Texas, Arizona, Georgia, South Carolina and Florida are among the 16 states with the highest degrees of poverty that have or have plans to implement high school graduation exams. For child poverty levels see 2001 Kids Count Data Online available: http://www.aecf.org/kidscount/kc2001/.

37. Ohanian, 1999.

38. Goodson & Foote, 2001.

39. McNeil, 2000.

40. Clarke, Haney, & Madaus, 2000.

41. The most influential research we found that substantiated the effectiveness of high-stakes testing policies came from Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S., 2000. Using NAEP data, researchers in this study recommended duplicating the high-stakes testing programs in North Carolina and Texas, although concrete evidence that high-stakes testing programs caused the achievement gains noted in those states was lacking. Only a few other studies have substantiated the positive effects of high-stakes testing. See Carnoy, Loeb, & Smith, 2000; Muller & Schiller, 2000; Scheurich, Skrla, & Johson, 2000; and Schiller & Muller, 2000.

42. Sacks, 1999 and Kohn, 2000b.

43. The attachment of accountability measures to high academic standards has enjoyed a full measure of bipartisan support for the last decade or more. Eilperin, 2001 and Valencia, Valenzuela, Sloan & Foley, 2001.

44. Haney, 2001; Haney, 2000; Neill & Gayler, 1999; and Sacks, 1999.

45. Firestone, Camilli, Yurecko, Monfils & Mayrowetz, 2000; Goodson & Foote, 2001; Haney, 2000; Heubert & Hauser, 1999; Klein, Hamilton, McCaffrey, & Stecher, 2000; Kohn, 2000a; Kossan & González, 2000; Kreitzer, Madaus, & Haney, 1989; McNeil, 2000; McNeil & Valenzuela, 2001; Reardon, 1996; Sacks, 1999; Thomas & Bainbridge, 2001; and Urdan & Paris, 1994.

46. Chiu, 2000.

47. Robelen, 2000.

48. Sacks, 1999.

49. Salzer, 2000.

50. Kossan, 2000.

51. Domench, 2000.

52. Gardner, 1999, p. 16.

53. Shorris, 2000; and Shorris, 1997.

54. "high-stakes tests," 2000.

55. Ibid.

56. Ibid.

57. Heubert & Hauser, 1999.

58. Linn, 2000, pg. 14.

59. Heubert & Hauser, 1999, pg. 75.

60. Heubert & Hauser, 1999.

61. McNeil & Valenzuela, 2001, pg..133.

62. McNeil & Valenzuela, 2001, pg. 134.

63. Wright, (Forthcoming).

64. This listing of "stakes" is not exhaustive. For example, local school districts and local schools may attach additional stakes to the consequences written into state test policies.

65. In 2004, grade promotion decisions in grades 3, 5, and 8 will be contingent upon student performance on Georgia's new Criterion-Referenced Competency Tests. Eventually, all Georgia students will have to pass promotion tests at each grade level.

66. Beginning in the fall of 2000, grade promotion became contingent on grades 4 and 8 performance on the new Louisiana Educational Assessment Program (LEAP 21) tests. Louisiana became the first state to retain students in grade using test scores.

67. Grade promotion in grade 8 is contingent on a combination of CTBS/5 test scores, student classroom performance, and classroom assessments.

68. A grade promotion "gateway" exists at grade 5. Beginning in 2002, promotion gateways will exist at grades 3 and 8.

69. Teachers in schools that perform poorly and are identified as low-performing by the state face the possibility of having to take a teacher competency test. Jones, Jones, Hardin, Chapman, Yarbrough, & Davis, 1999.

70. In 2002, promotion to the 5th grade will depend on a student's 4th grade reading score on the Ohio Proficiency Reading Test. Plans to make more grades promotion gateways are in progress.

71. In 2002, promotion to the 5th grade will depend on a student's 4th grade test scores. Plans to make more grades promotion gateways are in progress.

72. In 2003 students must pass the grade 3 reading test to be promoted to the 4th grade. In 2005 students must pass the grade 5 reading and math tests to be promoted to the 6th grade. In 2008 students must pass the grade 8 reading and math tests to be promoted to the 9th grade.

73. The state also uses student or school test results to evaluate teachers. Georgia has similar teacher accountability plans underway.

74. Quality Counts, 2001.

75. Information included has been pooled from the state department web sites and multiple telephone interviews with state testing personnel.

76. States with an asterisk (*) do not collect the percent of students who do not graduate or receive a regular high school diploma because they did not meet the graduation requirement at the state level. Almost 50% of the states do not collect this information. For these states, a rough percent was calculated by taking the number of 12th graders who on their last attempt before the intended graduation date did not meet the graduation requirement divided by the total 12th grade enrollment that year.

77. This percentage does not account for students who dropped out, who enrolled in alternative or GED programs, or who transferred out of state.

78. Mississippi does not collect the percent of students who do not graduate or receive a regular high school diploma because they did not meet the graduation requirement and does not collect any data on the test after the test is first administered. The number of 12th graders who took, failed, or passed the test was therefore unavailable so rough estimates were impossible to calculate.

79. New York's graduation exam requirement consists of a series of end-of-course exams. Students take the end-of-course tests as early as the 7th grade or when students complete each Regents course.

80. Students in North Carolina take the competency tests only if they did not pass a series of similar, end-of-grade tests taken at the end of the 8th grade.

81. Mehrens, 1998.

82. Neill & Gayler, 2001, pg. 108.

83. Fisher, 2000, pg. 2.

84. As cited in Haney, 2000.

85. Klein, Hamilton, McCaffrey & Stecher, 2000.

86. Schrag, 2000.

87. Heubert & Hauser, 1999, pg. 132.

88. NAEP data are not collected at the high school level by state. As such, NAEP scores will not be used as direct indicators of how high school students have been affected by high school graduation exams. However, if a state has a high school graduation exam in place it has appropriately been defined as one of the states with high-stakes written into K–12 testing policies. Accordingly, gains over time as compared to the nation will indicate how more general high-stakes testing policies have improved each state's system of education.

89. Judd, Smith, & Kidder, 1991 and Smith & Glass, 1987.

90. Fraenkel & Wallen, 2000; Glass, 1988; and Smith & Glass, 1987.

91. Information included was pooled from the state department web sites, multiple telephone interviews with state testing personnel, and Quality Counts, 2001. State testing personnel in all states but Florida and Virginia verified the information before it was included in this chart.

92. In 39% (7/18) of the high-stakes states – Georgia, Maryland, Mississippi, New York, South Carolina, Tennessee, and Virginia – students will take end-of-course exams instead of high school graduation, criterion-referenced tests once they complete courses such as Algebra I, English 1, Physical Science, etc. ... End-of-course exams seem to be the new fad, replacing high school graduation exams.

93. Since the 1960s student performance on New York's Regents Exams determined the type of diploma students receive at graduation – a Local Diploma or a Regents Diploma.

94. The competency tests are only given to 9th graders who did not pass the end-of-grade tests at the end of the 8th grade.

95. In 1983 students did not have to pass the Texas Assessment of Basic Skills to receive a high school diploma.

96. Glass, 1988 and Smith & Glass, 1987.

97. Glass, 1988, pg. 445–446.

98. Campbell & Stanley, 1963; Glass, 1988; and Smith & Glass, 1987.

99. From 1959 to 1989 the original version of the ACT was used. In 1989 an enhanced ACT was implemented but only scores back to 1986 were equated to keep scores consistent across time. This explains the slight jumps from 1985–1986 that will be apparent across all states. Although scores from 1980–1985 have not been equated, the correlation between scores from the original and enhanced ACT assessments is high: r =.96

100. See footnote #99 to explain the large increase illustrated from 1985–1986.

101. Smith & Glass, 1987.

102. ACT composite scores (1980–2000) were available on–line at http://www.act.org. or were obtained through personal communications with Jim Maxey, Assistant Vice President for Applied Research at ACT. We are indebted to him for providing us with these data.

103. SAT composite scores (1977–2000) were available on–line at http://www.collegeboard.com. or were provided by personnel at the College Board. We thank those at the College Board who helped us in our pursuit of these data.

104. Kohn, 2000a.

105. Trends were defined in the short term, as defined by the difference in score one year after the point of implementation, and in the long term, as defined by the difference in score the number of years from one point of intervention to the next or 2001 as compared to the nation.

106. Changes in participation rates as compared to the nation (1994–2001) are listed in parentheses.

107. Correlation coefficients represent the relationship between changes in score and changes in participation or exclusion rates for participating states with high-stakes tests. Only states with high-stakes tests were included in the calculations of correlation coefficients hereafter. Coefficients were calculated separately from one year to the next for the years for which data and participation rates were available.

108. These correlation coefficients were calculated using changes in score and changes in participation rates for the years in which data and participation rates were available.

109. Within states colleges may change their policies regarding which tests are required of enrolling students. This may affect participation and exclusion rates hereafter.

110. Changes in participation rates as compared to the nation (1991–1997 and 2000–2001) are listed in parentheses.

111. State NAEP composite scores (1990–2000) are available on–line at http://nces.ed.gov/nationsreportcard.

112. For more information on the NAEP, for example its design and methods of sampling see Johnson, 1992.

113. For further discussion see Neill & Gayler, 2001.

114. See Grissmer, Flanagan, Kawata, & Williamson, 2000 and Klein, Hamilton, McCaffrey & Stecher, 2000.

115. Johnson, 1992.

116. Neill & Gayler, 2001.

117. See the NAEP website at http://nces.ed.gov/nationsreportcard.

118. Haney, 2000.

119. This figure represents the r-square of each correlation coefficient that was calculated by squaring the correlations between change in score and change in exclusion rates year to year.

120. Changes in exclusion rates are listed next to changes in score hereafter. Scores and exclusion rates were

calculated as compared to the nation.

121. The exclusion rate for the nation in 1990 was not available. The exclusion rate was imputed by calculating the average exclusion rate for all states that participated in the 1990 8th grade math NAEP.

122. For a similar study see Camilli, 2000. Camilli tested claims made by Grissmer et al., 2000, that large NAEP gains made by students from 1992 to 1996 in Texas were due to high-stakes tests. Camilli found, however, that the cohort of Texas students who took the NAEP math as 4th graders in 1992 and then again as 8th graders in 1996 were just average in gains. Camilli analyzed cohort gains in Texas on the NAEP 1992 and 1996 math assessment only, however. This section of the study will expand on Camilli's work to include all states with high-stakes tests. Further, randomly sampled cohorts of students who took the NAEP math as 4th graders in 1996 and as 8th graders in 2000 and cohorts of students who took the NAEP reading as 4th graders in 1994 and as 8th graders in 1998 will be examined.

123. Toenjes, Dworkin, Lorence & Hill, 2000.

124. Klein, Hamilton, McCaffrey & Stecher, 2000.

125. AP data (1995–2000) were available in the AP National Summary Reports available on-line at http://www.collegeboard.org/ap.

126. Participation rates were calculated by dividing the number of AP exams that were taken by students in the 11th and 12th grade by each state's total 11th and 12th grade population. Grades received on the exams were calculated by dividing the number of students who received a grade of 3 or above, a grade of 3 being the minimum grade required to receive college credit, by the total number of 11th and 12th grade participants.

127. "Controlling" for participation rates was possible only in this analysis. Years for which we had participation rates matched the years for which we had the percentages of students who passed AP exams.

128. "Fisher," 2000.

129. "Advanced placement," 2000.

References

Administrative bonuses: Tied to student performance in Oakland. (2001, February 8). *The National Education Goals Panel Weekly* [On-line]. Available: http://www.negp.gov/weekly.htm

Advanced placement: Growth in Texas. (2000, August 31). *The National Education Goals Panel Weekly* [On-line]. Available: http://www.negp.gov/weekly.htm

Berliner, D. C. & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Bracey, G. W. (1995). Variance happens: Get over it! *Technos, 4* (3), 22–29.

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Education Policy Analysis Archives, 8* (42) [On-line]. Available: http://epaa.asu.edu/epaa/v8n42.html

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist, 30* (12), 1103-1126.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.) *Handbook of research on teaching*. Chicago, IL: Rand McNally & Company.

Carnoy, M., Loeb, S. & Smith, T.L. (2000, April). *Do higher test scores in Texas make for better high school outcomes?* Paper presented at the American Educational Research Association Annual Meeting, New Orleans, LA.

Chiu, L. (2000, October 3). Education issues fuel Capitol rally. *The Arizona Republic* [On-line]. Available: http://www.azcentral.com/news/education/1003AIMS03.html

Civil rights coalition sues over race discrimination in Michigan Merit Scholarship Program. American Civil Liberties Union Freedom Network, (2000, June 27). [On-line]. Available: http://www.aclu.org/news/2000/n062700a.html

Clarke, M., Haney, W., & Madaus, G. (2000). *High stakes testing and high school completion*. The National Board on Educational Testing and Public Policy [On-line]. Available: http://www.nbetpp.bc.edu/reports.html

Commission on Instructionally Supportive Testing. (2001). *Building tests to support instruction and accountability: A guide for policymakers* [On-line]. Available: http://www.aasa.org/issues_and_insights/assessment/

Domench, D. A. (2000, December). *School administrator web edition* [On-line]. Available: http://www.aasa.org/publications/sa/2000_12/domench.htm

Durbin, D. (2001, March 21). Merit awards fail minorities. *The Detroit News* [On-line]. Available: http://www.detnews.com/2001/schools/0103/21/c07d-202027.htm

Eilperin. J. (2001. May 23). House backs annual reading, math tests. *Washington Post* [On-line]. Available: http://washingtonpost.com/wp-dyn/education/A63017-2001May22.html

Elazar, D. J. (1984). *American federalism: A view from the states* (3rd ed.). New York: Harper & Row, Publishers.

Figlio, D. N. & Lucas, M. E. (2000). *What's in a grade? School report cards and house prices*. National Bureau of Economic Research [On-line]. Available: http://papers.nber.org/papers/w8019

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8* (35) [On-line]. Available: http://olam.ed.asu.edu/epaa/v8n35/

Fisher, F. (2000). Tall tales? Texas testing moves from the Pecos to Wobegon. Unpublished manuscript.

Folmar, K. (2001, August 9). Lawsuit delays payment of state bonuses to teachers. *San Jose Mercury News* [On-line]. Available: http://www0.mercurycenter.com/premium/local/docs/award09.htm

Fraenkel, J. R. & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). Boston, MA: McGraw Hill, Inc.

Gardner, H. (1999). *The disciplined mind: What all students should understand*. New York: Simon & Schuster.

Glass, G. V (1988). Quasi-experiments: The case of interrupted time series. In R. M. Jaeger (Ed.) *Complementary methods for research in education*. Washington, DC: American Educational Research Association.

Goodson, I. & Foote, M. (2001). Testing times: A school case study. *Education Policy Analysis Archives, 9* (2) [On-line]. Available: http://epaa.asu.edu/epaa/v9n2.html

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP test scores tell us*. Santa Monica, CA: RAND Corporation [On-line]. Available: http://www.rand.org/publications/MR/MR924/

Haladyna, T., Nolen, S. B. & Haas, N. S. (1991). Raising standardized test scores and the origins of test score pollution. *Educational Researcher, 20* (5), 2–7.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8* (41) [On-line]. Available: http://epaa.asu.edu/epaa/v8n41/

Haney, W. (2001). *Revisiting the myth of the Texas miracle in education: Lessons about dropout research and dropout prevention*. Paper prepared for the "Dropout Research: Accurate Counts and Positive Interventions" Conference sponsored by Achieve and the Harvard Civil Rights Project, Cambridge, MA [On-line]. Available: http://www.law.harvard.edu/groups/civilrights/publications/dropout/haney.pdf

Heller, D. E. (September 30, 1999). Misdirected money: Merit scholarships take resources from have-nots. *Detroit Free Press* [On-line]. Available: http://www.freep.com/voices/columnists/qehell30.htm

Heubert, J. P. & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press [On-line]. Available: http://www.nap.edu/html/highstakes/

High stakes tests: A harsh agenda for America's children. (2000, March 31). Remarks prepared for U.S. Senator Paul D. Wellstone. Teachers College, Columbia University [On-line]. Available:

http://www.senate.gov/~wellstone/columbia.htm

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29* (2), 95–110.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T. & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81* (3), 199–203.

Judd, C. M., Smith, E. R. & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Klein, S. P., Hamilton, L.S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8* (49) [On-line]. Available: http://epaa.asu.edu/epaa/v8n49/

Kohn, A. (2000a). *The case against standardized testing: Raising the scores, ruining the schools.* Portsmouth, N.H.: Heinemann.

Kohn, A. (2000b, September 27). Standardized testing and its victims. *Education Week* [On-line]. Available: http://www.edweek.org/ew/ewstory.cfm?slug=04kohn.h20

Kossan, P. & González, D. (2000, November 2). Minorities fail AIMS at high rate: Some backers talking change. *The Arizona Republic* [On-line]. Available: http://www.arizonarepublic.com/news/articles/1102aimsrace02.html

Kossan, P. (2000, November 27). By trying too much too quick, AIMS missed mark. *The Arizona Republic* [On-line]. Available: http://www.arizonarepublic.com/news/articles/1127aims27.html

Kreitzer, A. E., Madaus, G. F., & Haney, W. (1989). Competency testing and dropouts. In L. Weis, E. Farrar, & H. G. Petrie (Eds.) *Dropouts from school: Issues, dilemmas, and solutions.* Albany, NY: State University of New York Press.

Linn, R. L. (2000). Assessments and accountability. *Education Researcher, 29* (2), 4–15 [On-line]. Available: http://www.aera.net/pubs/er/arts/29-02/linn01.htm

Madaus, G. & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: Evidence from one hundred years of test data. In Orfield, G., & Kornhaber, M. L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press.

McNeil, L. M. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In Orfield, G., & Kornhaber, M.L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press.

McNeil, L. M. (2000). *Contradictions of school reform.* New York, NY: Routledge.

Mehrens, W.A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6* (13) [On-line]. Available: http://olam.ed.asu.edu/epaa/v6n13.html

Muller, C & Schiller, K.S. (2000). Leveling the playing field? Students' educational attainment and states' performance testing. *Sociology of Education, 73* (3), 196–218.

National Governor's Association. (2000, January 25). *1999 state initiatives on spending tobacco settlement revenues.* [On-line]. Available: http://www.nga.org/Pubs/IssueBriefs/2000/000125Tobacco.asp#Summary

Neill, M., & Gayler, K. (2001). Do high-stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation tests. In Orfield, G., & Kornhaber, M. L. (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press.

Neufeld, S. (2000, October 2). Backlash fermenting against school tests: Groups organize to complain about STAR. *San Jose Mercury News* [On-line]. Available: http://www.mercurycenter.com/premium/local/docs/backlash02.htm

Ohanian, S. (1999). *One size fits few: The folly of educational standards.* Portsmouth, NH: Heinemann.

Orfield, G., & Kornhaber, M. L. (Eds.) (2001). *Raising standards or raising barriers? Inequality and high-stakes*

testing in public education. New York: The Century Foundation Press.

Pallas, A.M., Natriello, G., & McDill, E.L. (1989). *The changing nature of the disadvantaged population: Current dimensions and future trends*. Center for Research on Elementary and Middle Schools. (ERIC Document Reproduction Service No. ED 320 655).

Paris, S. G. (2000). Trojan horse in the schoolyard: The hidden threats in high-stakes testing. *Issues in Education, 6* (1, 2), 1-16.

Quality Counts 2001. (2001). *Education Week* [On-line]. Available: http://www.edweek.org/sreports/qc01/

Reardon, S. F. (1996). *Eighth grade minimum competency testing and early high school dropout patterns*. Paper presented at the Annual Meeting of the American Educational Research Association (ERIC Document Reproduction Service No. ED 400 273).

Robelen, E. W. (2000, October 11). Parents seek civil rights probe of high-stakes tests in La. *Education Week* [On-line]. Available: http://www.edweek.org/ew/ewstory.cfm?slug=06ocr.h20

Ross, J. (2001, April 2). High-testing kids grab MEAP cash: More students savor $2,500 reward to use for college education, expenses. *Detroit Free Press* [On-line]. Available: http://www.freep.com/news/education/maward2_20010402.htm

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Books.

Salzer, J. (2001, April 29). Georgia dropout rate highest in nation: Years of reform achieve little, and some fear testing will only worsen problem. *Atlanta-Journal Constitution* [On-line]. Available: http://www.accessatlanta.com/partners/ajc/epaper/ editions/sunday/news_a3be6ad4f17dd10f007c.hml

Scheurich, J. J., Skrla, L & Johnson, J. F. (2000). Thinking carefully about equity and accountability. *Phi Delta Kappan, 82* (4), 293–299.

Schrag, P. (2000, January 3). Too good to be true. *The American Prospect* [On-line]. Available: http://www.prospect.org/archives/V11-4/schrag-p.html

Serow, R. C. (1984). Effects of minimum competency testing for minority students: A review of expectations and outcomes. *The Urban Review, 16* (2), 67-75.

Sheldon, K. M. & Biddle, B. J. (1998). Standards accountability and school reform: Perils and pitfalls. *Teachers College Record, 100* (1), 164-180.

Shorris, E. (1997). *New American Blues*. New York: W. W. Norton.

Shorris, E. (2000). *Riches for the poor*. New York: W. W. Norton.

Smith, M. L. & Glass, G.V (1987). *Research and evaluation in education and the social sciences*. Needham Heights, MA: Allyn and Bacon.

Swope, K. and Miner, B. (Eds.). (2000). *Failing our kids: Why the testing craze won't fix our schools*. Milwaukee, WI: Rethinking Schools, Ltd.

Thomas, M. D. & Bainbridge, W. L. (2001). "All children can learn:" Facts and fallacies. *Phi Delta Kappan, 82* (9), 660–662.

Toenjes, L. A., Dworkin, A. G., Lorence, J., & Hill, A. J. (2000). *The lone star gamble: High stakes testing, accountability, and student achievement in Texas and Houston* [On-line]. Available: http://www.brook.edu/gs/brown/bc_report/2000/Houston.PDF

U.S. Department of Education. (1983). *A nation at risk: The imperative for educational reform* [On-line]. Available: http://www.ed.gov/pubs/NatAtRisk/index.html

Urdan, T. C. & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy, 8* (2), 137-157.

Using tobacco settlement revenues for children's services: State opportunities and actions. (1999, October). National Conference of State Legislatures and The Finance Project. [On-line]. Available: http://www.financeproject.org/tobaccoattach.htm

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. E. (2001). Let's treat the cause, not the symptoms: Equity and accountability in Texas revisited. *Phi Delta Kappan, 83* (4), 318-321, 326.

Wright. W. E. (Forthcoming). The effects of high stakes testing on an inner-city elementary school: The curriculum, the teachers, and the English Language Learners. *Current Issues in Education* [On-line]. Available: http://cie.ed.asu.edu/

About the Authors

Audrey L. Amrein
College of Education
Arizona State University
Division of Educational Leadership and Policy Studies
PO Box 872411
Tempe, AZ 85287-2411

Email: amrein@asu.edu

Audrey L. Amrein is an Assistant Research Professional in the College of Education at Arizona State University in Tempe, Arizona. Her research interests include the study of large-scale educational policies and their effects on students from racial minority, language minority, and economically disadvantaged backgrounds. Specifically, she is interested in investigating the effects of high-stakes tests, bilingual education, and charter school policies as they pertain to issues of equity.

David C. Berliner
Regents' Professor of Education
College of Education
Arizona State University
Tempe, AZ 85287-2411

Email: berliner@asu.edu

David C. Berliner is Regents' Professor of Education at the College of Education of Arizona State University, in Tempe, AZ. He received his Ph.D. in 1968 from Stanford University in educational psychology, and has worked also at the University of Massachusetts, WestEd, and the University of Arizona. He has served as president of the American Educational Research Association (AERA), president of the Division of Educational Psychology of the American Psychological Association, and as a fellow of the Center for Advanced Study in the Behavioral Sciences and a member of the National Academy of Education. Berliner's publications include *The Manufactured Crisis*, Addison-Wesley, 1995 (with B.J. Biddle) and *The Handbook of Educational Psychology*, Macmillan, 1996 (Edited with R.C. Calfee). Special awards include the Research into Practice Award of AERA, the National Association of Secondary School Principals Distinguished Service Award, and the Medal of Honor from the University of Helsinki. His scholarly interests include research on teaching and education policy analysis.

Appendices

Appendices are available in either html format or Rich Text Format, the latter being a word processor file. The RTF file is 3.5 megabytes in size.

- Appendices in html.
- Appendices in Rich Text Format

BEST COPY AVAILABLE

Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

279

Education Policy Analysis Archives

Entrepreneurial Ambitions in the Public Sector: A Random Effects Model of the Emergence of Charter Schools in North Carolina

Linda A. Renzulli
University of Georgia

Abstract

In this article, I study charter schools as social innovations within the population of established public educational institutions. I begin by briefly outlining the history of public schools in the United States. Organizational theories are applied to explain the perpetuation of the structure of public schools since World War II. Next, I delineate the characteristics of educational reform movements in the United States by focusing on the charter school movement. Then, I use an evolutionary approach to study the environmental characteristics that drive the perceived need for innovation and the promotion of experimentation. Using data compiled from the North Carolina Department of Public Instruction, the Census Bureau, and North Carolina State Data Center, I examine the characteristics of the local environment that promotes the submission of charter school applications in North Carolina over a three-year period, 1996-1998. It is shown that school districts in need of school choice do have a higher mean charter school submission rate. Also, some community characteristics and available resources are important for the initial stage of charter school formation.

Introduction

Charter schools represent a new organizational form in the public school system. They are founded by teams of entrepreneurs that may include students, parents, educators, and community members. Instead of being part of the established bureaucratic structure, charter schools are independent of the rules and regulations of their associated school districts. In return for per-pupil expenditures and a release from some of the required bureaucratic structure, charter schools are held accountable by the state and local community for improved student achievement.

Since the first charter school was founded in Minnesota in 1992, the number of charter schools has grown at an increasing rate. In 1998, thirty-four states have charter school laws and twenty-six states plus the District of Columbia have operating charter schools (Center for Educational Reform 1999). As of October 1998, 1,128 charter schools

enrolled more than 250,000 children. These numbers are rising as new charters are accepted within states and as other states pass charter school legislation (Center for Educational Reform 1999).

The development of charter schools is the result of recent entrepreneurial activity and an example of how a new organizational form can arise as a result of a social innovation. In this paper, I study charter schools as an innovation within the population of established public educational institutions. I expect the founding of charter schools to exhibit distinct stages of innovation, creation, and maintenance, in which people have intentions to start a school, actually create a school, and finally, maintain a school. Specifically, I am interested in the environmental context in which these stages of social innovation unfold.

Evolutionary theory argues that organizational change in a population or community occurs as a result of three processes: variation, selection, retention and diffusion (Aldrich 1979). I treat the innovation of charter schools as an instance of a variation. First, variations in the environments can come from intentional or blind variation. Second, the variations must be developed and implemented or selected into the environment. And third, the innovation gains legitimacy and is retained through a struggle for resources. This paper will examine the first of the three phases— intentional creations (or at least attempted creations) of a new organizational form. Thus, the social innovation of charter schools within established educational institutions and structures is my primary concern.

I begin by briefly outlining the history of public schools in the United States. I apply organizational theories such as institutional theory to explain the perpetuation of the structure of public schools since World War II. Next, I delineate the characteristics of the educational reform movement in the United States by focusing on one of the many reform options—the charter school movement. Then, I use an evolutionary approach to study the environmental characteristics that insight community members and drive their perceived need for innovation and the promotion of experimentation.

Using data complied from the North Carolina Department of Public Instruction, the Census Bureau, and North Carolina State Data Center, I examine the characteristics of the local environment that cause charter school applications to be submitted in North Carolina over a three-year period. The analysis begins in 1996 when the North Carolina Sate Legislature passed the SL- 1997-430, which gave parents, teachers, and community members the legal option to publicly educate students through chartered schools. I use a Poisson random effects model to estimate the effects of community characteristics on the number of charter schools submitted for approval.

Traditional Public Schools

The Constitution of the United States does not explicitly say anything about public educational instruction; thus educational institutions fall under the jurisdiction of state and local governments. In the pre-industrial period (1607-1812) public schools were set up by boards of education and funded by taxes to educate children, usually poor children. (Note 1) Finally, the Tenth Amendment officially left the responsibility for education to the states, which reaffirmed the tradition of local control. Before World War II, public schools in the United Sates did not take on a clear institutionalized character, but after the war public schools became the norm (Hill, Pierce, and Guthrie 1997). Over time, public schools gained legitimacy and acceptance alongside private education. In fact, by 1950 a majority of American children attended public schools. More recent numbers show that public schools still educate most children. For example, in 1995, 33.9 million kindergarten through eighth grade children attended public school whereas only 4.427 million attended private schools (Hill 1995; Hill, Pierce and Guthrie 1997; United States Census Bureau 1998).

Public elementary and secondary education has been and still is a politically charged topic. States and local governments want more control of education, at the same time as the state and federal governments are ridiculed for not supporting education enough. As late as 1979, in response to the growing push for federal intervention, the 96th Congress passed Public law 96-88 that established a Department of Education. The Department of Education was instituted as a federal office to support more effective state and local educational institutions while still allowing state and local governments to maintain control of education. In fact, PL 96-88 clearly stated that the responsibility of education should remain in the hands of state and local systems. (Note 2) State and local boards of education determine curriculum, testing, and teaching in traditional public elementary and secondary schools, while the federal government provides special programs and funds to enhance and aid state controlled education.

Notwithstanding the freedom state boards of education have to create educational institutions and structures, public elementary and secondary schools have a remarkable resemblance to one other within and across state lines. Schools in Iowa have curriculum, structure, and schedule, similar to the schools in Maine. In fact, children can easily be moved from one state to the next because public education is so similar. The similarity between and within states can be explained if we consider schools an organizational form. Some organizational theories can explain the forces that produce similarities within a population, other organizational theories can explain why there are different organizational structures within the same population.

Schools as Organizations

There are a plethora of definitions for organizations in the literature. However, in my analysis I will use the following definition: "Organizations are goal-directed, boundary-maintaining activity systems" (Aldrich, 1979; Aldrich, 1999). I

have chosen this definition because it is broad enough to encompass public sector, non-business organizations such as schools. In addition, the core of this definition focuses on the social processes, initiation, and endurance of organizations. Thus, Aldrich's definition of organizations allows me to view schools, specifically charter schools, as organizations and focus on the process of their formation. I briefly explain the forces that produce similarity in organizational structures and then discuss the evolution of change via the educational reform movement and school choice.

Forces that Explain School Similarity

Institutional theory explains a great deal about the similarity among public schools (Meyer, Scott, and Deal, 1977). The system of public schools in the US maintains its legitimacy by conforming to an agreed upon set of rules and cultural expectations (Meyer and Rowan, 1977; Meyer and Rowan, 1978). Despite the variations in laws, schools almost universally educate children in similar subjects and similar ways, partly because of teachers' and administrators' sensitivity to public opinion (Bidwell, 1965). That is, the school as an organization faces formal pressures from state and local boards, informal pressures from parents and the community, and cultural expectations. Thus, schools are "highly penetrated organizations," sensitive to the environment (Meyer, Scott and Deal, 1977).

In organizational terms, schools are isomorphic, or constrained to resemble one another, due to the similar set of environmental conditions they encounter (DiMaggio and Powell, 1983). Schools do not, however, face competitive isomorphism because they are not in a population of "free and open competition" (DiMaggio and Powell, 1983; Hannan and Freeman, 1977). Traditional public schools face institutional isomorphism. In other words, schools fight to gain legitimacy for social acceptance. DiMaggio and Powell (1983) discussed three ways in which an organization undergoes isomorphism—coercive, normative, and mimetic. First, organizational leaders are coercively isomorphic because the organization needs to be political and social legitimized. Thus, an organization needs to follow both social and political norms. Second, normative isomorphism is related to professionalization and professional norms. That is, norms related to the development of the occupations that fill the organization. Finally, mimetic isomorphism is a result of uncertainty (see DiMaggio and Powell, 1983). I believe that coercive and normative mechanisms constrain the schools from deviating from the norm. (Note 3)

Generally, coercive isomorphism results from formal and informal political and social pressures on schools to meet cultural expectations (DiMaggio and Powell, 1983). Meyer and Rowan (1977) argued that organizations, which expand their presence in more than one social arena, increase their legitimacy by conforming to and creating institutionalized norms. State legislation, national test standards, and requirements for funding have helped perpetuate school systems and school structure. Educational institutions and their teachings enter students' and parents' lives through political debate, family issues, and socialization. Public education has been socially constructed to reflect the mainstream values and needs of students, parents, citizens, and teachers (Berger and Luckman, 1966).

Normative isomorphism is also an important factor in the perpetuation of the structure of public schools in the United States. Normative isomorphism stems from professional pressures on schools to conform to standards. Schools are created and run by people who have been selected from a larger population, formally educated, professionally socialized, and in many cases made official members of unions American Federation of Teachers (AFT) and professional organizations such as the National Education Association (NEA). Teachers are semi-professionals (Note 4) who are subject to normative pressures. Teacher colleges and teacher certification programs formally educate teachers. In their education, teachers are taught how to teach, what to teach, and how to behave professionally. In fact, students in schools of education have an apprentice-like program—student teaching—that gives them experience and time in a classroom. It is in the classroom that teachers learn professional standards and are expected to conform to them. For example, teacher colleges in New Jersey can prepare students to be teachers in any state. Therefore, professional standards created by teaching colleges and other associations perpetuate the structure of educational institutions and enhance similarity among schools across school systems. In sum, institutional norms and expectations have created traditional education in the U.S. Political structure, values, and professional norms play a part in creating public schools as we know them.

Furthermore, due to standardized norms and expectations, traditional schools are designed to serve the needs of the average child. Yet there are many children that do not fit the mold. Traditional schools often fail to meet the needs of students. However, every child must attend schools despite the goodness of fit between the student and the school. The only other option for students is private school, and most parents cannot afford private or religious schools. And thus, until recently, those children who could not afford alternatives to public education were stuck in traditional schools regardless of the school's ability to teach them. In the next section, I will use an evolutionary approach as applied to organizations to explain the reform movement in education. A caveat to keep in mind is that most of the work done on new organizational foundings and nascent entrepreneurship focuses on business organizations rather than public sector organizations. Nevertheless, I believe many of the same processes work for the evolution of the public sector.

Evolution of Schools

Because I am interested in the genesis of a new organizational form within the population of public schools, I must use a theoretical framework that is general enough to aid the understanding of social innovation. Institutional theory alone is not equipped to explain the recent changes in the educational system. (Note 5) Institutional arguments are

282

static; therefore, the theories they provide are not sufficient to explain the change in the nation's education system. Evolutionary theory, on the other hand, is a broad multi-dimensional theory that is most similar to ecology but uses principles form institutional theory, learning theory, cognitive theory, transaction cost economics, and resource dependence to explain changes in a population (Aldrich 1999). The theory directs our attention to the processes that produce patterned changes in a population (Aldrich, 1999).

Organizational ecological models help answer "why there are so many different kinds of organizations" (Hannan and Freeman, 1977). Until recently, researchers studying the public school system in the United States need not grapple with that question because most schools structures looked similar to one another. However, today with the emergence of "new" public educational structures, population ecology, and even more broadly evolutionary theory, is needed to help us understand why these new forms were created.

Changes in the population of schools can occur for various reasons. Research on foundings shows that a need for a change may be the impetuous for an innovation but is not sufficient or even necessary. For example, a school district may be falling behind national and state averages (their students are below the national or state average on standardized tests or are below grade level in math and reading), and yet an educational innovation never happens. On the other hand, school districts that produce above grade level students with above average and excellent test scores may produce many educational reforms. Although, the need for change may be an important part of the social innovation process, it can not explain all of it. Instead, social entrepreneurs are a key to the production of an innovation. Environments and communities may "breed" these entrepreneurs. Thus, the effects of community context are key to understanding social innovation in the public school system. I will model some of these community characteristics to predict social innovation.

Educational Reform

The educational reform movement focuses on school choice. Since the establishment of common schools in the mid-1800's, local or state boards of education have assigned children to schools. Thus, children were bound to a school by their geographical location. In contrast, the movement for school choice advocates free, public options for parents to send their children to schools other than those to which they are assigned. Alternative schools provide a choice in the public school sector rarely available before the reform movement. (Note 6)

The current school reform project has changed the nature and structure of public schools. Alternative schools are deliberate departures from the existing public educational form. Four major choice ideas have been experimented with in states around the country—magnet, voucher, contract, and charter schools. I briefly define the first three school choices and then spend the majority of time explaining charter schools.

Magnet schools allow students to go to a different public school than the one to which they are assigned. They typically have entrance exams (Nathan 1996). Vouchers are education "gift certificates" that allow children to use the money allocated to them for public schools, i.e. per-pupil expenditure, to go toward their schooling at a private institution. Vouchers have been rather controversial because public funds are often used for non-secular schools (see Nathan, 1996). Contract schools or school site management delegates the administrative and financial responsibilities of a school to a firm rather than the local government. In other words, business organizations are contracted by the local educational agency to run a school. Again, this option is controversial because for-profit organizations run the school and may have conflicts of interest between their profits and the well-being of the school.

Charter Schools

Charter schools represent an innovation in the population of education systems in the public sector. Their organization differs from a traditional public school, and their place in the district structure of education is different. Charter schools are public schools that break off from the school district's command, and yet they are still public schools. Charter schools do not need to comply with all of the district rules, and thus the bureaucracy usually associated with public education is reduced. In return, charters schools' missions and statements explicitly state that parents, teachers, and students will increase their participation in school-related activities (Bomotti, Ginsberg, and Cobb, 1999; Nathan, 1996). In addition, charter schools must be accountable for student achievement or they can be closed. Charter schools give freedom to teachers to use innovative techniques and to administrators to structure the school day to best suit the students they serve. Like other public schools, the only federal regulation with which charter schools must comply is that they must be non-sectarian and may not charge tuition (Koppich, 1997). Other regulations vary by state. For example, most states limit the number of charter schools that can be created, the number of non-certified teachers that can be hired, and the form of governance the schools can adopt. Nevertheless, the charter itself dictates the mission, curriculum, and population of the school. Thus, charter schools are public institutions, but they differ in many ways from traditional public schools.

Charters are funded by state allotment for each student, but lack the capital that public schools get from the state and local community. When families place their child in a charter school, the state and local funds allotted to that child are moved from the traditional public school and given to the charter school. It is easiest to picture children sitting in a classroom with dollar signs over their heads. When they move from a traditional public school to a charter school the dollar sign moves with them, but the desk and building they were sitting in not. In other words, the only money each child gets is the dollar amount assigned to the individual, but not the capital used to build the traditional school.

Funding is probably the most debated and controversial issue of charter schools. Per-pupil expenditure pays for teachers, books, and supplies. Opponents of charter schools argue that losing one child to a charter school means the loss of between 3 to 5 thousand dollars, but fixed costs for the public school, such as the number of teachers or their salary, remain the same. Thus, reallocating per-pupil expenditures to go to charter schools rather than the fixed costs in the traditional public school will hurt public schools systems (Koppich, 1997). In fact, opponents in Ohio, Illinois, and Oregon have used this argument to defeat charter school legislation (Marks, 1998).

However, researchers in North Carolina have shown that the negative financial consequences incurred by districts are overstated. First, very few students per district leave the traditional public schools (Hassel, 1998). Thus, the financial burden is minimal for most districts. Second, districts can contract services to charter schools for fees to recoup some of the money they lost to charter schools. In other words, some of the per-pupil expenditures lost to charter schools can be gained back through contracting educational and transportation services (Coulson, 1996). Finally, the financial loss felt by districts may be an added incentive for traditional public schools to improve their accountability to families. In fact, charter school advocates intended for charter schools to create competition for students and decrease the degree of monopoly traditional public school have over education (Hassel, 1998). In the aggregate, charter schools receive less than what traditional public schools receive from public funds and do not put a great burden on their feeding school district (Note 7) (Hassel, 1998). Some even go as far as to say that charter schools do not receive their "fair share" of public funds for a public schooling (Hassel, 1998).

Charter Schools and Small Businesses

Business organizations and charters schools exhibit an important similarity. The similarity is important for understanding charter schools via organizational theories and using an evolutionary approach. Both charter schools and businesses must show gains. Like all organizations, a charter school must define and meet its goals to succeed and survive. Achievement gains in a charter school are equivalent to profit gains in a business. Charter schools must be accountable for student achievement. If they fail to do so, a charter school will lose its charter just as if a business fails to meet its profit goals it will fail. In addition, charter schools, like businesses must stay out of debt. For example, in Chapel Hill/Carborro North Carolina, a charter school closed mid-year early in 1999 because of a $50,000 debt they could not repay. Charter schools and businesses alike cannot run in the red and survive.

But charter schools are not like small business in many significant ways. Charter schools must comply with state regulation and criteria. For example, if a group of people want to start a charter school, they must first write an application and submit it to either the local educational agency or the state board of education for preliminary approval. (Note 8) Once the charter obtains preliminary approval, it is then sent to the State Board for final approval. If the charter is not approved, many states have an appeal process. In any case, the State Board reviews the charters to make sure they match the criteria laid out by the state legislation.

A small business does not go through such an application process. Under most circumstances, almost anyone can open a business without approval from the state or local government. Furthermore, after a given period of time, the state board reviews the existing charter schools. Small businesses are not subject to official review. Finally there is a difference in the number of organizations legally allowed to form. State regulations limit the competition for charter schools, while other business organizations are in a more or less free and open market competition. Most states have a limit on the number of charter schools. For example, North Carolina has a cap of 100 schools but Mississippi has a cap of 6 charters. However, the Department of Commerce does not limit the number of donut shops or delis allowed to open in a state. Thus the survival of a charter school depends less on the density of the population of charter schools then does the survival of business organizations and so I expect different factors to affect schools than businesses such as community involvement (see Hannan and Freeman 1989 for a complete discussion of density dependence).

Charter Schools as a Variation in the Population of Public Educational Institutions

Variations occur as responses to need, resource availability and mobilization when people actively attempt to generate alternatives to existing forms. Variations are important for creating competition, an implicit part of the evolution of organizations. They are also important because variations help generate differences within a population.

Charter schools are intentional variations that arise from innovative and experimental groups of initiators often to solve the problem of complacent and monopolistic schools (Coulson, 1996). Public schools have had an almost complete exclusive jurisdiction of the education of American youth (besides religious and private schools) and thus the population of schools has had very little variation and competition (Scott, 1992). In fact, traditional public education meant that a school would be run by the local board where choice in education meant that one would have to pay tuition to a private institution (Hill, 1995; Hill, Pierce and Guthrie, 1997). Because state and local government had a monopoly on education, public elementary and secondary schools did not have to be accountable for growth and gains in student achievement.

However, state and local governments, those who have monopolistic control over education, are the only ones who can make school choice a legal option. Thus, the system, rather than the environment, controls innovations in education at large. In fact, the initiation of the variation can only occur in states where the legislature has passed a

charter school law. Someone must bring an initial bill to the state legislature. These policy entrepreneurs (Mintron 1997) help create a formal arena for educational entrepreneurs to experiment with a new educational form. Without the law, variations in public education are impossible.

Charter Schools in North Carolina

North Carolina passed legislation in 1996, four years after Minnesota passed the first charter school law, that allows public funds to go to charter schools. According to The Center for Educational Reform, North Carolina has a strong charter school law (see Appendix 1 for details of North Carolina Charter School regulations). "A strong law (also known as a "live," "effective," "expansive" or "progressive" law) is one that fosters the development of numerous, genuinely independent charter schools" (Center for Educational Reform, 1999).

North Carolina had 63 operating charter schools as of November 1998. The schools are located in rural areas in the western counties, cities such as Raleigh and Charlotte, and affluent villages such as Chapel Hill/Carrboro. The schools serve a range of students including average students, at risk students, special needs children, and exceptional children among other populations. Controversy exists over the quality of education found in charter schools, the populations they serve, the number of students they teach, and the teachers teaching them, (for example Jackson, 1998).

Data

The data for this analysis come from several sources. First, I collected the population of submitted charter school applications to the North Carolina State Board over a three-year period from the first legal year of charter schools in 1996 to 1998. These data will be used to create the dependent variable in the analysis: number of submissions of charter school applications to the State Board of Education. These data are from the Department of Public Instruction—Office of Charter Schools Recommendations for Preliminary and Final Approval summary reports. There were 184 charters submitted between 1996 and 1998, of which 66 were accepted. (Note 9) I was able to obtain the year the charter was submitted, acceptance status, grade level, county or local educational agency in which the school would be located, and charter type. For this analysis, I will only use the county and year for which the charter was submitted.

The second source of data is from the 1996 USA Counties DataBase (from Census Bureau) and North Carolina State Data Center (SDC). The USA Counties DataBase is a conglomeration of data compiled from the 1990 census and Current Population Surveys (1992-1995). SDC is a consortium of state and local agencies that provides data and information about North Carolina. I use the 1995 data. Because local educational agencies or LEAs (Note 10) and counties have an almost 1 to 1 correspondence in North Carolina (Note 11), I can use information about LEA and counties as contextual-level variables.

Finally, the third source of data comes from NC Department of Public Instruction. In 1997, NC implemented the ABC's, a program to monitor all public schools in North Carolina, in accordance with the School-Based Management and Accountability Act (1996). (Note 12) School growth and performance is measured by composite scores computed for expected growth/gain, exemplary growth/gain, and a percentage of students at or above grade level. Excellent, distinguished, progressing and low performing schools are identified based on the composite scores.

The data are in a stacked or LEA-year format. In other words, each of the 118 LEAs have 3 observations, one corresponding to 1996, one to 1997, and one to 1998. Therefore, I have 354 observations of LEA years. I then use a year dummy variable to capture period effects. However, the other independent variables for the county characteristics of the LEAs are time invariant. In other words, I continue to use the same values of the independent variables over time. I use lagged county data (community characteristics prior to 1996) to reflect the county atmosphere prior to submission of a charter school application.

Hypotheses

I propose three main hypotheses regarding the effect of the environment on charter school application submissions.

Sources of Dissatisfaction: Charters are educational options. If schools in a district are meeting the public expectations for educating its student body, the need for school choice may be attenuated. Conversely, those districts that are not meeting the educational expectations of parents, teachers, and administrators may generate increased dissatisfaction with public schools. Therefore, I expect that quality of LEA schools will be inversely related to the number of charter school submissions because people will respond to their dissatisfaction. In addition, LEAs with few school -age children as a proportion of the total residence will be less likely to need charter schools. For example, charter schools are often aimed at special needs children. The number of special needs children in LEAs with relatively few school-age children will be fewer than in LEAs with a greater proportion of school-aged children.

H1: LEAs that have a higher percentage of low performing schools will have a greater number of submitted charter school applications.

H1a: LEAs with a smaller proportion of school age children will have fewer submitted charter school applications.

Resources: School funding is an important part of the charter school application process. Those people who submit an application must show that the charter is financially viable. According to the state law, charter schools receive state and local per-pupil expenditures for each student that attends the school but do not receive capital or startup funds. Therefore, the more per-pupil funding, the easier it may be to start and maintain a school and thus the more likely one would be to submit an application.

H2: The greater the state and local per-pupil expenditure in an LEA, the greater the number of submitted charter school applications.

Characteristics that promote experimentation: The community characteristics may have a positive or negative effect on charter school submissions. For example, as stated above, strong institutional norms may inhibit experimentation and social innovation. These norms may be produced by coercive, normative, or mimetic isomorphism. Therefore, the extent to which a community has institutional norms may inhibit or enable social innovation.

Despite the possible institutional constraints on social innovation, new organizations even in the nascent stages need founders. The need for a new form is not sufficient to understand the solutions (Aldrich, 1999). Therefore, information about the types and groups of people in a community may be helpful to understand who starts a new form. Active and innovative groups are likely to see a need and act on it to create a new charter school. Therefore, I include information about the political environment as an indicator of the activity of people in the community. In addition to the voter population, the political affiliations of voters may be important. Although nation wide both Democrats and Republicans have supported charter schools, Democrats seem more likely to support educational initiatives such as charter schools. Therefore, counties that are more Democratic will be more likely to start a charter school.

Furthermore, having a college degree is a personal characteristic that may increase the likelihood that a person would initiate a charter school. College educated parents may be more active in their children's education, question educational practices, and have the skills to start a new educational form.

H3a: The greater proportion of registered voters, the more likely there will be a greater number of submitted charter school applications.

H3b: The greater proportion of voters who are Democratic, the more likely there will be a greater number of submitted charter school applications.

H3c: The greater proportion of college educated people in the county the more likely there will be a greater number of submitted charter school applications.

Variables

Dependent Variable: Number of submitted charter school applications.

I use the number of applications rather than the number of accepted charters to avoid testing the political structure of the approval process. I am interested in the effects of the environment on the decision to initiate a new educational form in a LEA. Figure 1 shows an approximate Poisson distribution of the percent of LEA's applications submitted to the state from 1996 to 1998. A LEA may have submitted an application in 1996 but not in 1998. Therefore, the percentage of LEAs is not mutually exclusive from one year to the next. There were 57 local educational agencies (49 percent of the 118 LEAs) that never submitted a charter school application between the years of 1996 and 1998.

BEST COPY AVAILABLE
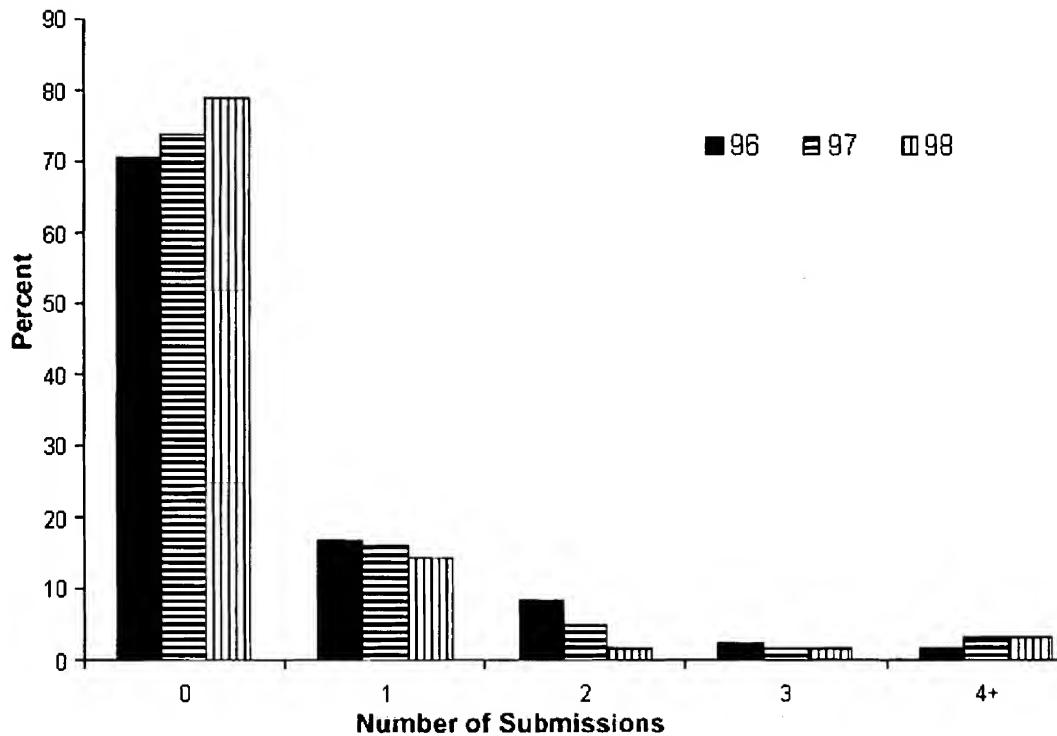
Figure 1. Percent of LEAs with Charter School Application Submissions by Year

Table 1
Independent Variables

Year (for period effect)	Dummy for 1996, 1997, 1998. (1996 = reference category)
LEA Public Schools (Based on the ABC's 1996-1997):	
High Performing schools	A ratio of the number of schools identified by the state as being a school of distinction or excellence over the number of public schools in the LEA. Schools of distinction meet their expected growth or have at least 80 percent of their students performing at or above grade level. An excellence school has at least 90 percent of their students performing at or above grade level. (range 0 to 100 percent of the schools in the LEA)
Low Performing Schools	The number of schools that fail to meet their expected growth or have less than 50 percent of the students perform at or above grade level over the number of schools in the LEA. (range 0 to 100 percent of the schools in the LEA)
School aged children	Proportion of school aged children in the county, 1995 (age 5-18)
Characteristics to Promote Experimentation:	
Voter registration	The proportion of people (16+) registered to vote, 1995
Political affiliation	The proportion of registered voters who are Democrats, 1995
Percent College Graduates	The proportion of people in the county who are have received a college degree, 1990
Funding:	
State Per-pupil expenditure	The dollar amount of money allotted to each child in the LEA by the state divided by 100 in 1996
Controls—Demographic Indicators:	
Racial	Proportion white in the county, 1995

composition of the county	
Density per square mile	Proxy for rural or urban area is the number of people per square mile in the county, 1995.
Unemployment Rate:	The unemployment rate in the county in 1994.
Per capita income	Per capita income in 1995

Control variables

I have included variables to control several other community characteristics. Many charter schools are found in cities rather than small affluent suburbs (though exceptions exist), therefore, I control for density, unemployment, per capita income, and racial composition as urban proxies.

Method and Model

The Poisson Model

Count data are discrete, non-negative integers that enumerate the number of events. Count dependent variables cannot be treated as continuous in a linear regression model or as binary in a logit model because the estimates will be inefficient, inconsistent, and biased (Long, 1997). A linear OLS model will not produce reliable estimates because the count variable does not have a normal distribution. The log of a count variable cannot correct the problem because the log of a zero value is undefined. Therefore, we need a count model to estimate the effects of an environment on the number of events occurring.

Count models can correct for the problems of the OLS by transforming the dependent variable. The Poisson regression model (PRM) determines the probability of the count of the event occurring by using the Poisson distribution. Poisson is the simplest of the count models due to the properties of the probability density function [$f(x)$= ($^x e^{-8}/ x!$] but it is also restrictive. One of the assumptions of the PRM is that the dependent variable's mean and variance are equal. Thus, the distribution assumes that the event count is time independent where the conditional mean of the error is 0 and the errors are heteroscedastic since var(e|x)= E(y|x)=exp(xb). In these data, the mean is .5 and the variance is 1.7; thus the data are not distributed in an exact poison distribution.

Random Intercept model

A simple Poisson model is not appropriate for correlated data. In such instances, the maximum likelihood estimation of a Poisson model will result in biased and inconsistent estimates. Because of the longitudinal data structure in this analysis and the use of time invariant covariates, the most appropriate model is the random effects model for a count dependent variable. The data for this analysis are clustered by school systems over three years. To correct for the clustering and account for the distribution of the dependent variable a Count Random Effects Model is most appropriate. (Note 13) The "extgee" command in Stata© (Note 14) can estimate a Poisson Random Intercept model that corrects for over-dispersed dependent variables.

BEST COPY AVAILABLE

Results

The exploratory analysis of the dependent variable is shown in Table 2. As shown, the mean number of submissions per LEA per year is .5 with a variance of 1.7. The correlations between the other independent variables are all below .6. The number of submissions is correlated with the independent variables (results not shown here). The significant correlations gave me enough confidence to proceed with the multivariate PREM.

Table 2
Poisson Random Effects Model of
Number of Submitted School Applications

Variables	Coef.		Std. Err.	Odds Ratio
Intercept	4.9300		2.60	--
YEAR				
1997	0.0000		0.18	1.000
1998	-0.2231		0.18	0.800
NEED (Hypothisis 1)				
% Low Performing LEA Schools	7.2200	+++	2.26	1336.440
% High Performing LEA Schools	-0.0186		0.18	0.982
% School Age	-1.6124		6.17	0.199
RESOURCES (Hypothesis 2)				
State funding 96	-0.1600	+++	0.05	0.850
EXPERIMENTERS (Hypothsis 3)				
% Registered to Vote	-2.7795	+	1.50	0.062
Democrat	0.0012	^	0.00	1.001
% College Grad	5.9474	+++	1.78	382.739
CONTROLS				
Unemployment Rate	0.1149		0.09	1.122
Density per Sq. Mile	0.0001		0.00	1.000
Per capita Income	0.0001		0.00	1.000
% white	-0.1823		0.69	0.833
ICC or Rho			0.106	
chi2(df=15)			280.37 ***	
Number of observations			354	
Number of Clusters			118	
Number of obs per cluster			3	

+p< .05, ++p< .01, +++p<.001, one-tailed test
^p<.1, *p<.05, **p< .01, ***p<.001, two-tailed test

A series of nested models was run, systematically including each set of independent variables. By comparing the chi-squares and degrees of freedom, the most restrictive model was chosen as the final model. The results of the nested models are not shown here.

Effects of environment on submission

As predicted in hypothesis 1, a greater proportion of low performing schools does increase the mean number of charter school applications. As shown in Table 2, a 1 percent increase in low performing schools in the LEA increases the mean number of submissions by 134 percent. Hypothesis 2 was not supported. The direction of the effect of state funding is negative but not significant. Surprisingly, a 100 dollar increase in state funding decreases the mean number of submissions by 15 percent. One explanation for reversal in the predicted direction is that school districts that are getting more state money may be better off than those districts with less money and thus do not need school choice. This finding would be in direct contradiction to The Coleman study (1966) but may support educators such as Kozol who believe that the inequality in per-pupil expenditures does inhibit learning and performance for some students (Kozol, 1991). Further research should explore the causes and consequences of per-pupil expenditures on school performance and alternative schooling.

I show some support for the third set of hypotheses that environmental characteristics promote experimentation through active groups of people. An increased proportion of registered voters actually decreased the mean number of submission in a county. The variable I use for voters may be too passive a measure to capture the activity of community members. Voter registration does not necessarily mean that people are actually voting. I have used this measure but the reversed sign indicates to me that a better measure, such as voter turnout, would be more appropriate. Counties that have a greater proportion who are Democratic voters have more submitted charter school applications over the three years in North Carolina. For ten percent increase in proportion of Democratic affiliated voters there is a 1.2 percent increase in the mean number of submitted applications net of all other variables. In addition, a one-

percent increase in the proportion of college grads in a county increases the mean number of submissions by 38 percent. The effect is so large due to the distribution of the variable. A transformation of the variable such as taking the natural log may help reduce the effect but transformations make interpretation more difficult to interpret intuitively.

In sum, counties with more democrats and college educated people and a greater percent of low performing schools will tend to have more submitted charter school applications. Further research should examine the characteristics of the individuals who initiate charter schools in their communities. It is interesting that the control variables I specified did not have a significant effect on the submission of charter school applications. Thus the racial composition, unemployment, and density of an area do not encourage or dissuade charter school development. This may be an area for further research.

Conclusion

The first part of the process of creating a new organizational form was analyzed here. The environmental context explained what types of communities are most likely to initiate a new educational form. LEAs in need of school choice (defined by the percentage of low performing schools) in fact do have a higher mean submission rate. It has also been shown that some characteristics about the community and available resources are important for the initial stage of charter school formation.

This analysis has shown that the environment is an important factor in predicting nascent school foundings. It is the first step in our understanding of charter schools as a new educational organizational form. I would like to use more direct measures of community and professional activity to test the assumption that certain groups will be social entrepreneurs that initiate new organizational forms. Thus, information about the entrepreneurs and the resources they access will further illuminate our understanding of charter schools as a new organizational form.

There is now evidence that school districts that fail to meet high performance standards have a higher mean charter school submission rate than districts that have high student outcomes. This suggests that charter schools are being developed in areas where there is a perceived need for them. The question remains if the students who attend charter schools are the underserved in those districts, which would include students at-risk and students of color. Recent studies in Arizona have shown that charter schools, in fact, do more to segregate districts and schools than to integrate them (Cobb and Glass, 1999). Do my findings here about the initiation of charter schools and the findings about charter school racial composition suggest that charter schools are being used as a means of white flight without residential mobility?

Some community characteristics and available resources are important for the initial stage of charter school formation. This research suggests that policy makers interested in encouraging charter school formation should provide financial resources to nascent founders for their endeavors but should also ensure that all students get a chance at attending charter schools.

Appendix I

Charter School Legislation: Profile of North Carolina's Charter School Law North Carolina Law passed 1996, amended 1997.	
Number of Schools Allowed	100 (maximum 5 per district per year)
Number of Charters Operating	59
Additional Schools Approved (Oct 1998)	5
Approval Process Eligible Chartering Authorities	Local school boards, state board of education, University of North Carolina institution, local board and UNC approval subject to final approval by state board
Eligible Applicants	Person, group of persons, or non-profit Corporation
Types of Charter Schools	Converted public and private, new starts (but not home-based schools)
Appeals Process	Charter denied by local school board or UNC institution may be appealed to state board of education
Formal Evidence of Local Support Required	For conversions, majority of teachers and uncertified staff at school must support; evidence that a significant number of parents support conversion must also be provided; districts must provide and sponsors must consider impact statement
Recipient of Charter	Applicant
Term of Initial Charter	Up to 5 years

Automatic Waiver from Most State and District Education Laws	Yes from state laws; yes from district regulations except for local-board-sponsored charters, which must negotiate with sponsor district for waivers from district rules
Legal Autonomy	Yes
Governance	Specified in charter
Charter School Governing Body Subject to Open Meeting Laws	Yes
Charter School May be Managed or Operated by a For-Profit Organization	Charters may not be granted directly to for-profit organizations, but charter schools may contract with for-profit organizations to run the school
Transportation for Students	Charter schools must provide same transportation assistance as district public schools
Facilities Assistance	Districts required to lease available public space to charters so long as it is "economically viable;" charters may lease space from sectarian organizations so long as sectarian symbols are removed
Technical Assistance	Department of education must provide technical assistance to charter school applicants upon request
Reporting Requirements	Charter school must comply with reporting requirements established by state board of education in the Uniform Education Reporting System; charter school must prepare annual report for chartering authority and state board; state board must prepare annual report on academic progress, best practices, and effect of charter schools on districts for legislature
Funding Amount	100 percent of state and district operations funding follows students, based on average district per-pupil revenue; special needs funding also follows the student
Path	State funds flow directly to charter school; local funds pass through district to charter school
Fiscal Autonomy	Yes
Start-up Funds	No state funding; federal charter school funding will be applied to start-up costs
Teachers	
Collective Bargaining / District Work Rules	For charter school sponsored by local school board teachers remain subject to district work rules unless they negotiate to work independently; for all other charter schools, teachers are not subject to district work rules (North Carolina is a right-to-work state)
Certification	25 percent of teachers in elementary charter schools and 50 percent in secondary charter schools may be uncertified
Leave of Absence from District	Up to 6 years
Retirement Benefits	Teachers qualify for state retirement plan during leave of absence from district; state has defined charter employees as public employees and has asked the IRS for a formal ruling on providing retirement benefits to teachers not on leave from a district
Students	
Eligible Students	All students in state
Preference for Enrollment	Children of charter's professional staff; in a charter's first year of operation the lesser of 10 percent or 20 slots may be reserved for children of founding board members; for public conversions, students in attendance area of former public school (for private conversions, students attending the school prior to conversion may not receive preference)
Enrollment Requirements	Not permitted
Selection Method (in case of over-enrollment)	Lottery
At-Risk Provisions	Preference in the approval process is given to charter schools designed to serve at-risk students
Racial Balance Provisions	After one year, charter school must reasonably reflect racial balance of district (or, if serving special population, must resemble the balance of that population in the district)
Mandated Assessments	Student assessments required by state board of Education

Other Features	
School Size	Charter schools must have a minimum number of students (65) and teachers (3), though exceptions are allowed; may increase by 10 percent without additional approval from sponsor
Termination of Charter	If two thirds of teachers and support staff request, charter may be terminated

Source: The Center for Education Reform (Fall, 1998 http://edreform.com/laws/NorthCarolina.htm)

Notes

1. At the same time, religious schools continued to educate many of the children between the ages of 5 and 17. In addition, exclusive private schools were formed by the upper class to separate privileged children from the lower and middle class or, for that matter, the children of the Nouveau riche (Levine 1980).

2. The Constitution only specifies that public education must comply with the separation of church and state.

3. Mimetic mechanisms are probably not the cause of isomorphism in the school population because uncertainty is rare.

4. The professional status of teachers has been debated; nevertheless, teacher colleges and unions are at the very least fighting for the professionalization of the occupation.

5. Institutional theories can explain why the diffusion process is slow for educational reform in curriculum and organizational development, but the diffusion process is beyond the scope of this paper. For more information about the diffusion (Mort 1958; Owens 1981: p. 237-249).

6. Private, and religious schools also provide variations in educational institutions but not in the public realm. New York's magnet schools are also an example of an educational choice but competitive based on entrance exams.

7. The data for financial burden come from an analysis of funding in North Carolina. Depending on state law, some states may find more or less burden to their districts.

8. This varies state by state. In North Carolina, charter applications may be sent to either the LEA or the state for preliminary approval.

9. Three of the 66 accepted schools were closed sometime after operation.

10. Local Educational Agency, school system, and school district are all synonymous.

11. There are 100 counties and 118 Local Educational Agencies in NC.

12. The results are publicly available and can be found in "A Report Card for the ABCs of Public Education, Volume I."

13. A suitable model would be a negative binomial REM to account for the over-dispersion in the dependent variable. In other words, the variance is not equal to the mean and thus the dependent variable actually has a negative binomial distribution. Using a negative binomial model would relax the assumption of an equal mean and variance of the Poisson distribution.

14. Copyright 1999. Stata Corporation, College Station, Texas.

References

Aldrich, Howard E. (1999). *Organizations Evolving*. London: SAGE.

Aldrich, Howard E. (1979). *Organizations and Environments*. Englewood Cliffs, NJ: Prentice-Hall.

Berger, P. and Luckman, T. (1966). *The Social Construction of Reality*. Garden City, NY: Doubleday.

Bidwell, Charles. 1965. The school as a formal organization. Pp. 972-1022 in *Handbook of Organizations*, edited by

J. G. March. Chicago, Il: Rand McNally.

Bomotti, Sally, Rick Ginsberg, and Brian Cobb. (1999). Teachers in charter schools and traditional schools: A comparative study." *Education Policy Analysis Archives, 7*(22). Available online: http://epaa.asu.edu/epaa/v7n22.html.

Center for Educational Reform. (1999). Answers to Frequently Asked Questions About Charter Schools. , vol. 2000: http://edreform.com/school_reform_faq/charter_schools.htm.

Cobb, Casey D. and Glass, Gene V. (1999). Ethnic segregation in Arizona charter schools. *Education Policy Analysis Archives, 7*(1). Available online: http://epaa.asu.edu/epaa/v7n1/.

Coleman, J., Campbell, E., Mood,A., E. Weinfeld, E., Hobson, D. York, R. and McPartland, J.. (1966). *Equality of Educational Opportunity.* Government Printing Office, Washington, DC.

Coulson, Andrew. (1996). Markets versus monopolies in education: The historical evidence." *Education Policy Analysis Archives, 4*(9). Available online: http://epaa.asu.edu/epaa/v4n9.html.

DiMaggio, Paul J. and Powell, Walter W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review 48,* 147-160.

Hannan, Michael T. and Freeman, John H. (1989). *Organizational Ecology.* Cambridge, MA: Harvard University Press.

Hannan, Michael T. and Freeman, John H. (1977). The population ecology of organizations. *American Journal of Sociology 82,* 929-64.

Hassel, Bryan. (1998). Charter School Funding in North Carolina. Durham, NC: North Carolina Charter School Resource Center.

Hill, Paul T. (1995). Reinventing Public Education. Santa Monica, CA.: RAND.

Hill, Paul T., Pierce,Lawrence C. and Guthrie, James W. (1997). *Reinventing Public Education: How Contracting Can Transform America's Schools.* Chicago: The University of Chicago Press.

Jackson, Deidra. (1998). Charter Schools Touted for Minorities. in *The News and Observer.* Charlotte.

Koppich, Julia E. (1997). Considering Nontraditional Alternatives: Charters, Private Contracts, and Vouchers. *The Future of Children 7,* 96-111.

Kozol, Jonathan. (1991). *Savage Inequalities: Children in America's Schools.* New York, NY: Harper Perennial.

Levine, A. (1980). The rise of American boarding schools and the development of a national upper class. *Social Problems 28,* 63-94.

Long, J. Scott. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks: SAGE Publications.

Marks, Tiayana. (1998). Arguments and Groups Affecting Current Attempts to Pass Charter School Legislation. Available online: http://epaa.asu.edu/eppa/v3n13/legis.html.

Meyer, John, W., Scott, Richard and Deal, Terrence. (1977). Research on school and district organization, in *The Structure of Educational Systems: Explorations in the Theory of Loosely Coupled Organizations,* edited by M. R. Davis, T. Deal, J. Meyer, B. Rowan, W. R. Scott, and E. A. Stackhouse. Stanford: Stanford Center for Research and Development in Teaching.

Meyer, John W. and Rowan, Brian. (1977). Institutionalized organizations: Formal structure as myth and ceremony." *American Journal of Sociology 83,* 340-63.

Meyer, John W. and Rowan, Brian. (1978). The Structure of Educational Organizations. Pp. 78-109 in *Environments and Organizations,* edited by M. Meyer, J. Freeman, M. T. Hannan, J. Meyer, W. G. Ouchi, J. Pfeffer, and W. R. Scott. San Francisco, CA: Jossey-Bass Publishers.

Mintron, Michael. (1997). Policy entrepreneurs and the diffusion of innovation. *American Journal of Political Science* *41*, 738-770.

Mort, Paul. (1958). Educational Adaptability. Pp. 32-33 in *Administration for Adaptability*, edited by D. H. Ross. New York: Metropolitan School Study Council.

Nathan, Joe. (1996). *Charter Schools: Creating Hope and Opportunity for American Education*. San Francisco, CA: Jossey-Bass Publishers.

Owens, Robert G. (1981). *Organizational Behavior in Education*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Scott, W. Richard. (1992). *Organizations: Rational, Natural, and Open Systems*. Englewood Cliffs: Prentice Hall.

United States Census Bureau. (1998). Statistical Abstract of the United States: 1998. , Washington, DC.

About the Author

Linda Renzulli
Assistant Professor
University of Georgia
Department of Sociology
113 Baldwin Hall
Athens, GA 30602

Phone: 706-542-3213
fax: 706-542-4230
Homepage: http://www.uga.edu/soc/renzulli.htm
Email: renzulli@uga.edu

Linda A. Renzulli is an Assistant Professor in sociology. Her research interests include organizations and education. She is currently working on a several charter school projects including racial composition of teachers in charter schools and the initiation of charter school legislation and charter schools across the United States.

EPAA Editorial Board

BEST COPY AVAILABLE

Education Policy Analysis Archives

The Amalgamation of Chinese Higher Education Institutions

David Y. Chen
Huazhong University of Science & Technology

Abstract
The 1990s witnessed revolutionary change in China's higher education system, particularly through radical mergers. The reform process and its background are detailed here, with a case study focusing on Zhejiang University. After nearly 15 years of painstaking effort, the reform goals for the higher education system have been met, and a decentralized, two-tiered administrative system has been installed. However, the most hotly debated reform has been the amalgamation of universities. The need to optimize China's system of higher education has a background dating back about 50 years, when the first reordering of higher education took place. The reordering and its results are described, and the causes and after effects of this reform are detailed.

Never before has Chinese higher education undergone such momentous changes, and never before has higher education attracted so much attention from both the general public and authorities at all levels. A new awakening has been brought about in higher education and as a result of this new leap forward. As the vice-premier of the Chinese government announced on August 24, 2000, at a meeting of Congress, China's optimization of the administrative structure of higher education has been basically and successfully fulfilled (Li, 2000).

The main target of reform was to change the obsolete system under which universities were owned and run by a variety of central industry ministries, in order to establish a fairly decentralized, two-tiered management system. In this system, administrative powers would be shared by both central and local governments, but with the local governments being required to play a major role. After nearly fifteen years of painstaking effort, this two-tiered administrative system has been finally installed.

During the whole process of reformation, the guidelines were *gongjian*(joint administration), *tiaozheng*(adjustment), *hezuo*(cooperation) and *hebing*(merger). *Gongjian*, or joint administration between the central government and local levels illustrates the potential of provincial governments in the construction of universities. *Tiaozheng*, or adjustment, calls for a shift in the balance of administrative power from the central government to local levels. *Hezuo*, or cooperation, requires universities in the same area to cooperate by making full use of resources owned by different institutions. Now, 452 institutions have changed their masters, and only a few more than 100 universities still remain directly under the administration of the central government. Seventy-one flagship universities are under the jurisdiction of the Ministry of Education (MOE), and another fifty or so professional institutions (e.g., defense, sports, civil aviation, etc.) are temporally under those corresponding ministries. *Hebing*, or merger, refers to the attempt to

merge several universities and colleges into one. Although the amalgamation of universities and colleges is the most difficult decision to make, nevertheless a total of 612 higher education institutions have been merged into 250 (Li, 2000), though these mergers have sometimes been perfunctory and unpleasant.

The Process

The process of reforming the administrative system of higher education can be divided into three stages.

1. *The brewing stage* (1985 to 1992). In 1985, the central authority declared the first act to restructure higher education. New ideas were widely publicized, reform was encouraged, and although, sporadic pilot experiments were indeed performed, no substantial progress was made. Still, the necessary foundation for further change had been laid.

2. *The exploration stage* (1992 to 1997). By 1992, the State Commission of Education (now MOE) actively sought a solution to the problem of segmentation between horizontal (called "bars") and vertical (called "blocks") departments, and by tentatively moving some institutions from the control of central ministries to provincial governments. In 1992, Guangdong province pioneered the pilot reform by co-constructing Zhongshan University and the Huanan University of Science and Technology under an agreement with the State Commission of Education. The administration of the Guangzhou University of Foreign Languages was also moved from the State Commission of Education to Guangdong province. Meanwhile, mergers between universities were used as a mechanism to change the structure of higher education. The Tianjing College of Foreign Trade, owned by the Ministry of Foreign Trade, was transferred and at the same time amalgamated into Nankai University. During this period, some large-scale universities were established through amalgamation. In May of 1992, seven colleges in the city of Yangzhou in Jiangsu Province (Jiansu Agriculture College, Yangzhou Teacher's College, Yangzhou Technical College, Yangzhou Medical College, Jiangsu Business College, and Jiangsu College of Water Conservation) were merged into a single new institution, Yangzhou University. Yangzhou University thereafter covered a wide range of disciplines, and as a result, became then the most comprehensive and perhaps the largest university established since the 1950s. However, the most tortured merger was between Sichuan University and the Chengdu University of Science and Technology in April, 1994. This was the very first case of amalgamation between strong universities. In the reordering of the 1950s, these two universities split from the then original Sichuan University. In fact only one road cuts the campus in two. However, after decades of development, they were almost equally strong, though both suffered the deficits of provincialism and restrictions brought about by the arrangement of narrowly set disciplines. Both were later voluntarily incorporated into one institution with formal support from the State Commission of Education. In addition, other comprehensive and large-scale universities were also created by combining several institutions. These include Nanchang University in Jiangxi province, Yanbian University in Jilin province, Shanghai University, Qingdao University in Shangdong province. By 1998, 207 institutions had been merged into 84 (Bao, 1998).

3. *The full-scale advancement stage* (1998 to 2000). In 1988, an important meeting was held in Yangzhou, Jiangsu province to speed up the reform of the higher education administrative system. At the same time, the fourth campaign of governmental restructuring was officially unveiled in the central government. Its goal was to change the role of government in the market economy emphasizing more macro-regulation rather than unnecessarily detailed micro-direction. As a result, the number of departments of the State Council was reduced from 40 to 29 (GUO, Nei, 2000), and the size of governmental staffs was reduced by half. Professional ministries were no longer permitted to run higher education institutions. Instead, universities and colleges were required to separate from their originally affiliated departments and find their own means of survival. Some were to be decentralized to the localities, others were to be transferred to the Ministry of Education, mainly by merging with those universities that were already under the direct administration of the Ministry of Education. In this stage, 1,232 institutions were radically changed through decentralization and amalgamation. About 406 universities have been restructured into 171 since 1996 (Ji, 2000). Consequently, the amalgamation of universities and colleges was accelerated. Before 2000, the focus was on the readjustment of administrative powers of those universities, which were separated from their former masters. However, from the start of 2000, a general advancement was pushed forward. In just six months, 778 institutions affiliated with 49 departments under the State Council had been restructured.

The entire process rested on two basic premises. First, all top-rate universities should be comprehensive, should include most disciplines, and should be big enough to handle large enrollments. Secondly, most medical universities should be incorporated into comprehensive educational institutions, and recognized as essential parts of first-class universities.

There are two kinds of merger. One is to merge closely located institutions sharing the same or similar disciplines, but affiliated with different governmental departments. This is done in order to increase efficiency and effectiveness, and to tackle the problem of segmentation and provincialism. Another is to form many larger and stronger universities by combining leading universities with relatively narrow disciplines. This is done in order to build representative and supposedly world-class universities. As a result, a number of bigger and stronger universities emerged with comprehensive fields of study in literature, arts, science, technology, agriculture and medicine. For example, Tsinghua University, China's leading university in science and engineering, incorporated the Central Academy of Arts, a leading institute in art design. The new Zhejiang University, the new Wuhan University, and the new Huazhong University of Science & Technology were each created from four smaller universities, and the new Jilin University was created through the merger of five smaller universities. The latter, which now consists of five campuses, presently has the largest student enrollment in China consisting of about 46,000 full-time resident students,

130 undergraduate programs, and 180 postgraduate programs including 71 doctoral programs (Chen, 2000). The new Zhejiang University covers all disciplines except military science, has five campuses, 40,000 full-time students, a staff of ten thousand, 98 undergraduate programs, 193 postgraduate programs, and 106 doctoral programs. Established in 1988 by merger of Zhejiang University, Hangzhou University, Zhejiang University of Agriculture, and Zhejiang University of Medical Science, it is one of the largest and most comprehensive universities in today's China (Wen & Bi, 2000).

Most strikingly, the majority of strong medical universities have been absorbed into flagship universities in this large-scale merger. Beijing University took in Beijing University of Medical Sciences, the best in its field. Shanghai No.1 University of Medical Sciences, one of the best, was incorporated into Fudan University. Other medical school mergers include Tongji University of Medical Science and Huazhong University of Science and Technology, Hunan University of Medical Sciences and Zhongnan University, Huaxi University of Medical Science and Sichuan University, Hubei University of Medical Science and Wuhan University, Zhejiang University of Medical Science and Zhejiang University, Baiqiouen University of Medical Science and Jilin University, and Xi'an University of Medical Sciences and Xi'an Jitong University. Many ambitious universities dreaming of becoming so-called world-class institutions are finding ways to incorporate with the left over medical universities to avoid being perceived as inferior to others in competition for resources and status in the hierarchy of higher education.

Nevertheless, the new round of amalgamations of universities and colleges was eventually completed, having proceeded reluctantly for some universities and willingly for others, but all reacting to the polices of the central government. For more detail on these amalgamations, please see the Appendix: Major Mergers of Universities Currently Under the Direct Administration of the Ministry of Education.

Behind the Amalgamation

Why did China's system of higher education need to be optimized? The reason can be found in an examination the situation about fifty years ago when the first reordering of higher education took place.

When the People's Republic of China was set up in October 1949, the higher education sector was fairly small. Among the 205 higher education institutions at that time, 60 percent were publicly-owned, 40 percent were privately-owned or owned by foreign missionary organizations, and enrolled in total just 117,000 students (only 2.2 students per 10,000 population), and 16,000 teachers (MOE, 1984). In 1951, after about two years of minor readjustments, among the 211 universities and colleges there were 49 universities that had at least three schools or departments of discipline classes; 91 independent colleges that had only one or two schools or departments of discipline classes; and 71 special higher institutions that in general covered only one or two disciplines. However, when the large-scale industrial construction of the First Five-year Plan began nationally, such a system of higher education revealed very distinct drawbacks. Geographically, most higher institutions were located in coastal areas. In 1949, while 79 of the 205 were in Beijing, Shanghai, Jiangsu, and Guangdong provinces, there were only nine in the large northwestern areas. In the structure of disciplines, there were too many arts and literature, social sciences and humanities programs on campuses, but little engineering, agriculture and animal husbandry, medical sciences, and teacher training programs. There were a hundred institutions that offered programs in politics and law, and seventy offered programs in economics and finance. Students studying engineering, agriculture and animal husbandry, and medical science accounted for a mere 31.5% of total enrollments (Liu, 1991).

As required under the First Five-year Plan, large-scale economic restructuring and construction concentrated on a series of industrial projects with the support of the then Soviet Union. As socialist construction needed a large pool of labor talent, mainly technical professionals, a major reorganization of higher education became inevitable. However, what pattern would be followed: the traditional Chinese pattern, the communist revolutionary pattern, or some foreign pattern?

At this point, however, the international political climate suddenly changed. The intensification of the Cold War forced the newly established China to close its doors to the West, and moreover, China's participation in the Korean War from 1950 to 1953 led Chinese politicians to a closer relationship with the socialist Soviet Union. Politically, economically, and culturally, the Chinese government chose an all-out emulation of Soviet Union patterns and practices, with the cordial assistance of large numbers of Soviet experts both as consultants to the various ministries, and as teachers and researchers in a number of specific institutions. Therefore, higher education increasingly assumed a Soviet Union character.

The first large-scale reform of higher education was put into practice in 1952 and 1953 under full guidance from the Soviet Union. This program was called *yuanxi tiaozheng*, which in Chinese means the reordering of colleges and departments. The reordering involved two important aspects: the geographical rationalization of the higher education layout, and the reestablishment of new types of institutions with special emphasis on the development of new engineering universities, both polytechnical and specialized, and teachers colleges. The primary concern was to restructure the whole higher education system in ways which would immediately serve the economic and political objectives set by the First Five-year Plan. Each institution and each program had a specially designated mission oriented directly to an industrial sector or a specific product or technical process. Consequently, all institutions were put under scrutiny and reorganized by department and specialization. Tactically, universities that had spent decades developing fairly comprehensive programs of literature and the arts, sciences, engineering, agriculture, law and

medicine were destroyed in order to build new specialist universities, colleges, and departments. All related departments, programs, teachers, equipment, and books in the related higher education institutions were concentrated and moved to one newly designed institution so as to build a specialized college. Almost overnight specialist colleges mushroomed across the nation. In order to ensure an even geographical distribution of each type of higher institution, six major regions (the Northwest, the Southwest, the Central South, the East China, the North China, and the Northeast) were designated as the basic units for political-administrative planning. Each region was allowed to establish one or two comprehensive universities (i.e., liberal arts and/or science(s) institutions), one or two polytechnical universities or colleges, one major teachers college, one to three agriculture universities or colleges, and other specialized institutions.

Following the two years of reform from 1951 to 1953, the total number of higher institutions decreased from 211 to 182. Among the 182 institutions, there were 14 comprehensive, 39 engineering, 31 teachers, 29 agricultural, 29 medical, 6 financial, 4 political and law, 8 language, 15 art, 5 sports, 2 ethnic, and 1 other (CIES, 1984). While the Ministry of Higher Education (now MOE) had been the only legitimate administrative organ for higher education, and directly administered comprehensive, polytechnical and key teachers colleges, specialized institutions were rationed to and administered by the corresponding central specialized ministries, (e.g., all mechanical institutions were under the direct leadership of the Ministry of Mechanics, all agricultural institutions under the Ministry of Agriculture, etc.). The whole process was, to a large degree, centrally planned and monitored. The only institutions administered at the provincial level were small local teachers colleges. In order to improve the geographical balance, from 1955 to 1957 a small-scale restructuring was initiated by moving five coastal universities to the hinterland, and building twelve new institutions there. Although other reforms were tried in the 1960s and 1970s, the overall structure and framework remained relatively unchanged after the radical reordering of the 1950s.

This system had two obvious characteristics. From the perspective of the administrative structure, professional ministries owned and administered relevant specialized institutions. The so-called *bumen banxue* (institutions owned and operated by ministries) led to compartmentalization, insularity, and self-protection in each sector, and an almost-closed system of higher education. All programs were set according to the sector's needs; all students were recruited on the basis of the sector's needs. In other words, all resources of specialized institutions in a certain system belonged to the affiliated ministry. Of course, such a system gave incentives for every ministry to support its own institutions both financially and politically, and to develop its own *zhuanye* (majors or specialized fields) and employ its own graduates. Naturally, institutions in such a closed system had no need to worry about their survival. Under a system of highly centralized planning, such closed systems were somewhat appropriate to the needs of the fledgling economy and social development. However, as the prevailing policy was turning from highly centralized planning to a market-oriented economy, such a pattern was no longer rational. Institutions oriented to self-aggrandizement in a closed system resulted in a great waste of scarce resources and inefficiency. In 1998, for example, 147 four-year universities and colleges had on average fewer than 2,000 students on campus, a figure representing 24.9% of all four-year institutions. The enrollment in each of the 108 two-year and three-year specialized institutions was below 1,000 students, accounting for 25.5% of this category. Improving efficiency and effectiveness became the biggest motivation for the full-scale amalgamation of institutions.

From the perspective of the functional type of institutions, all universities and colleges had become too narrow and specialized in disciplines, with engineering, agriculture, medicine, etc., artificially separated from liberal arts and basic sciences. As a result, there were no genuine comprehensive universities. This fragmentation of disciplines runs counter to the current trend of scientific integration, and of course, is detrimental to the cultivation of a body of students with broad vision and an integrated structure of knowledge. Thus, in the 1990s there was a cry from both within and outside for the establishment of several truly comprehensive universities with enough strength for competition in the world market. This is another important reason for the large-scale amalgamation of higher education institutions.

Still these reasons are not sufficient to explain the large-scale amalgamation of institutions. The most important external force came from the fourth governmental restructuring initiated in 1998. Through this restructuring, all national ministries were optimized and minimized. Except for very special and national security related universities, no one was permitted to remain under the leadership of the central ministries except the Ministry of Education. Those universities originally attached to the specialized ministries had to find ways to survive whether through decentralization to the provincial governments, being moved to the Ministry of Education, or through merging. Thus was the push towards large-scale amalgamation of universities finally accelerated.

Disquiet During the Amalgamation

Opponents have argued that radical amalgamation is full of risk, especially when it involves those institutions that are forced or are at least reluctant to be combined. The act of merger, these opponents argue, does not always raise the quality of a university, but in fact, might even dampen the enthusiasm of those institutions merged. Instead of radical amalgamation, some have pointed to other ways of improving efficiency, including internal restructuring of disciplines and increasing enrollment. Another criticism is that the existing 1,000-plus general institutions cannot meet the education needs of a country with 1.3 billion people, so to reduce this small number through merger is in fact not necessary.

Mergers between bigger and stronger universities can result in difficulties caused by the fusion of campus cultures,

personnel, disciplines, and the pressure of management of large-scale universities. Many oppose these mergers between the bigger and stronger universities, but support it in the case smaller and weaker institutions, and also approve of the annexation of smaller and weaker institutions by bigger and stronger universities because of the relative ease with which the former can be manipulated and managed. Because of this opposition, the central government has attempted to enhance its administration and encouraged mergers through financial subsidies. In any event, the period of rampant amalgamation of higher educational institutions in China is over. Now is the time for reflection and facing new challenges of institutional management. Whether amalgamation will be regarded as a success or not, only history will tell.

A Case in Point: Zhejiang University

Zhejiang University was founded in 1894 as Qioushi Academy in Hangzhou City, Zhejiang province. By 1950, Zhejiang University had earned a national and international reputation, and had become one of China's best and most comprehensive universities. The university had 24 departments in 7 schools: the school of literature, the school of sciences, the school of engineering, the school of agriculture, the teachers college, the school of law, and the school of medicine. In addition there were ten institutes, affiliated hospitals, factories, farms, and a forestry center.

However, when the reordering of institutions and departments began in 1952, Zhejiang was changed from a comprehensive university to a polytechnic institute. Then it was divided into some specialized colleges, and certain parts were moved to other universities. The school of medicine was incorporated with another medical college as an independent Zhejiang College (renamed University in the 1990s) of Medical Science. Its school of agriculture also became another unattached Zhejiang College (also renamed University in the 1990s) of Agriculture, and its teachers college was merged with another university thereby forming a new school first known as Zhejiang Teachers College, but later named Hangzhou University. And the major part of its school of sciences was transferred to Fudan University that had been designated as a comprehensive university. The department of forestry was transferred to Northeast College of Forestry in Harbin, Helongjiang province, and the department of animal husbandry and veterinarian medicine was transferred to Nanjing College of Agriculture. The department of aeronautics was shifted to Nanjing College of aeronautics, and department of water conservancy was transferred to East China College of Water Conservancy in Nanjing, Jiangsu province. Some of its teachers were ordered to four other universities. After this unprecedented reordering, the new Zhejiang University had only four departments: mechanics, chemical engineering, civil engineering, and electrical mechanics—a true polytechnic university.

Then in 1988, another revolutionary readjustment began which essentially reversed the reordering of 1952. Zhejiang University, Hangzhou University, Zhejiang University of Agriculture, Zhejiang University of Medical Science (four universities that had the same ancestor) were amalgamated into a new Zhejiang University. The new Zhejiang University, which today is the most comprehensive university in China, boasts disciplines ranging from philosophy and the sciences to agriculture and management, and a student population second only to Jilin University in enrollment. In all it has 20 schools, 70 departments, 183 institutes, more than 40,000 students on five campuses, and a staff of almost 30,000.

Conclusion

The massive amalgamation of China's higher education system is basically concluded. The reform reflects the revolutionary changes in Chinese society, and general developmental trends in higher education from around the world.

References

Bao, Daosu and Xu, Guangming. (1998). Nanchang University. *China Education Daily*, May 2,1998.p.2.

Chen, Fanbo. (2000). Five Universities Merged into Jilin University. *China Education Daily*, June 13, 2000. P.1

CIES (Central Institute of Educational Science). (1984). *Education Memorabilia of the People's Republic of China (1949-1982)*. Beijing: Education Science Press, 1984. P.70-71, Item NO.5372.

Guo, Nei. (2000). Government reform makes progress. *China Daily*. Oct.10, 2000.

Ji, Baocheng. (2000). Farewell to the Administration by Central Ministries. *China Education Daily*, September 25, 2000. P.1.

Li, Lanqing. (2000). Working Report on Implementing the Strategy of Rejuvenating the Nation through Science and Technology Delivered at the Seventeenth Session of Standing Committee of the Ninth National People's Congress on August 24, 2000.

Liu, Yifan. (1991). Higher Education History of Contemporary China. Wuhan: HUST Press.

MOE (Ministry of Education of China). (1984). The Statistics of Educational Achievements of China(1949–1983). Beijing: Renming Education Press.

Wen, Hongyan and Bi, Quanzhong. (2000). Work Hard to Create Word First-class Zhejiang University and Its Road of Amalgamation and Reform. *Renming Daily*. May 12,2000.

About the Author

David Y. Chen
Institute of Higher Education
Huazhong University of Science & Technology
Wuhan, China 430074; e-mail: davidcyc@263.net

David Y.Chen (Chinese Name: Chen Yunchao) is currently a PhD candidate at the Institute of Higher Education of Huazhong University of Science and Technology in China and a lecturer at the University of Logistical Engineering. His research interests include the university presidency, management of higher education, and international comparative research. David can be contacted at davidcyc@263.net.

Appendix
Major Mergers of Universities Currently Under the
Direct Administration of the Ministry of Education

University	Institutions Merged
Beijing University	Beijing University, Beijing University of Medical Sciences
Tsinghua University	Tsinghua University, Central Academy of Techniques Arts
Nankai University	Nankai University, Tianjing College of Foreign Trade
Northeast University	Northeast University, Gold college
Jilin University	Jilin University, Jilin Industry University, Baiqiouen University of Medical Sciences, Changchun University of Science and Technology, Changchun College of Postal and Communication
Fudan University	Fudan University, Shanghai University of Medical Sciences
Tongji University	Tongji University, Shanghai Railway University, Shanghai College of City Construction, Shanghai College of Construction Materials
Shanghai Jiaotong University	Shanghai Jiaotong University, Shanghai Agriculture College
Huadong University of Science and Technology	Huadong University of Science and Technology, Jinshan Petrochemical College
Donghua University	China Textile University, Shanghai Textile College
East-China Teachers University	East-China Teacher's University, Shanghai College of Education, Shanghai No.2 College of Education, Shanghai Teacher's College for Children
Dongnan University	Dongnan University, Nanjing College Railway Medial Medical Sciences, Nanjing Jiaotong College
Hefei Industry University	Hefei Industry University, Anfei College of Technology
Zhejiang University	Zhejiang University, Hangzhou University, Zhejiang University of Medical Sciences, Zhejiang Agriculture University
Shangdong University	Shangdong University, Shanghai University of Medical Sciences, Shanghai Industry University
Wuhan University	Wuhan University, Wuhan University of Hydroelectric, Wuhan University of Mapping and Survey, Hubei University of Medial Sciences
Huazhong University of Science and Technology	Huazhong University of Science and Technology, Tongji University of Medial Sciences, Wuhan College of City Construction, Wuhan Training College of Science and Technology for Cadres

Wuhan University of Science and Technology	Wuhan Industry University, Wuhan University of Auto-Industry, Wuhan University of Communication Technology
Hunan University	Hunan University, Hunan University of Finance
Zhongnan University	Zhongnan Industry University, Hunan University of Medical Sciences, Changsha Railway College, Changsha Industry College
Zhongshan University	Zhongshan University, Zhongshan University of Medical Sciences
Sichuan University	Sichuan University, Chengdu University of Science and Technology, Huaxi University of Medical Sciences
Chongqing University	Chongqing University, Chongqing University of Construction, Chongqing College of Construction
Xi'an Jiaotong University	Xi'an Jiaotong University, Xi'an University of Medical Sciences, Shannxi College of Finance
Northwest University of Agriculture and Forestry Sciences	Northwest University of Agriculture Sciences, Northwest College of Forestry Sciences, Institute of Water Conservancy of China Academy of Sciences, Northwest Institute of Irrigation works of the Ministry of Water Conservancy, Shannxi Academy of Agriculture, Shannxi Academy of Forestry, Northwest Institute of Plants of China Academy of Sciences
North Jiaotong University	North Jiaotong University, Beijing College of Electric Power
Beijing University of Chinese Medicines	Beijing University of Chinese Medicines, Beijing College of Acupuncture and Bone Injury
University of Foreign Trade and Economy	University of Foreign Trade and Economy, China College of Finance
Zhongnan University of Finance and Law	Zhongnan University of Finance, Zhongnan University of Law
Chang'an University	Xi'an Road Transportation University, Northwest College of Construction Engineering, Xi'an College of Technology

BEST COPY AVAILABLE

other vols. | abstracts | editors | board | submit | book reviews | subscribe | search

303

The Possibility of Reform:
Micropolitics in Higher Education

Susan Haag
Arizona State University

Mary Lee Smith
Arizona State University

Abstract
The purpose of this case study was to examine the restructuring of an institution of higher education's teacher preparation program and to assess the possibility for systemic reform. Although teacher education represents a vital link in not only the educational system but in curricular reform, the increased expectations for educational reform made this institution unavoidably more political. These conditions meant that the study of micropolitics was critical to understanding how organizations change or fail to initiate change. Any effort to reform an organization requires examination of the reform effort's underlying assumptions, social and historical context for the reform, and how reform is congruent with the values, ideologies, and goals of the constituents. This case will serve those critiquing reform and also takes the extant K-12 micropolitical research into the heretofore unstudied realm of higher education therefore impacting reform at the post secondary level. Schools are vulnerable to a host of powerful external and internal forces. They exist in a vortex of government mandates, social and economic pressures, and conflicting ideologies associated with administrators, faculty, and students. Efforts to reform school are confounded by competing political agendas. At the very least, reform is an opportunity for political action by people in power. While literature regarding effective schools touts strong leadership and shared values, accomplishing school reform continues to remain problematic. Despite the widespread interest and infusion of resources for restructuring teacher education, the history of educational reform shows that initiatives have often failed. The study began with the micropolitical hypothesis that the educational system comprises diverse constituencies with differing ideologies regarding schooling. Qualitative methodology was employed to portray intra-organizational processes, to provide concrete depiction of detail, and to study social change. Micropolitics and symbolic interactionism, the theoretical frameworks for the study, influenced the design and production of research and functioned as the interpretive focus. The study followed a multi method approach to understand meanings in context and to interpret these patterns in light of broader contexts. We employed the following multiple methods to generate a credible account of constituent ideologies: 23 semi-structured interviews, document review, and observational data. Data reveal fundamental differences in the images of five constituencies in these areas: curriculum, teachers, pupils, and teacher education and support the micropolitical assertion that systemic reform is unobtainable.

The Present Study

This present study examines competing views of institutions and their reform by the empirical study of a single case, the efforts of a College of Education (COE) to reform its undergraduate teacher preparation program. This case of local reform at a large urban university takes place within a broader context of nation-wide reform and restructuring. Teacher education is a primary concern in current restructuring efforts because it is a vital link in the educational system. However, despite the widespread interest and infusion of resources for restructuring teacher education, the history of educational reform shows that initiatives have often failed. Though administrators have often interpreted poor outcomes as evidence that individuals (i.e., teachers) fail to comply with reform agendas, evidence from this and earlier cases suggests that intra-organizational processes reflect micropolitical phenomena, not a lack of teachers' professional integrity (Ball, 1987; Blase, 1991; Noblit, Berry, & Dempsey, 1991). The major focus of this study is the impact of micropolitics on the possibility and success of the reform initiative in higher education. From the micropolitical perspective, reformers and researchers alike must examine internal processes that facilitate or impede change. Thus, our primary goal was to achieve a deeper understanding of the ideologies, goals, and values of each major teacher education constituency in regard to curriculum, teachers, pupils, and teacher education.

The Case

Why this case? We used this case because teacher preparation is an essential link in the educational system and reform, yet the COE Teacher Preparation program was in crisis due to reform efforts. The increased expectations for reform rendered this institution and its communities more political (West, 1999). Through the lens of micropolitics, we examine these conflicting views of institutions and their reform in order to understand how organizations change. To accomplish this goal, we focus primarily on the culture within the COE, that is, on intra-organizational processes.

As Ragin (1992) pointed out, "It is impossible to do research in a conceptual vacuum because the empirical world is limitless in detail and complexity. We make sense of its infinity by limiting it with our ideas" (p. 217). Evidence and ideas are mutually dependent; we transform evidence into results with the aid of ideas and make sense of theoretical ideas by linking them to empirical evidence. Cases are not empirical units of theoretical categories, but are the products of basic research operations. A case provides texture to the space between theory and empirical evidence. This case study employs an interpretive approach to qualitative methodology using micropolitics and symbolic interactionism as conceptual frameworks.

To address research issues, we examined the efforts of a major university's College of Education to reform its Teacher Preparation Program (TP, for short). The first year of the TP program was designed to enable students to plan, implement, and evaluate instructional activities in a variety of disciplines. TP students spent at least four hours a week in a school environment the first semester and six hours a week during the second and third semesters. Students enrolled in methods classes specialized in elementary, early childhood, secondary, special, or bilingual education during the third semester. They participated in a required field experience each semester of the program prior to student teaching, which was completed during the fourth semester of the program. The teacher education program graduated approximately 500 students a year, each with a certification in elementary, secondary, and special education.

The study originated during the attempted restructuring of the teacher preparation program in the College of Education. Efforts to reinvent teacher education began by involving external and internal communities. The dean scheduled meetings, focus groups, and retreats to identify discrepancies between current teacher preparation and desired practices. He called for efforts to strengthen preservice teachers' pedagogical knowledge, to increase collaboration between the university and local schools, and to educate preservice teachers to better serve the needs of an increasingly culturally diverse student population.

During COE meetings, faculty, staff, and students associated with teacher education were invited to participate in the creation of a shared vision. Faculty and staff were involved in the process to discuss concerns and to propose reform efforts; however, different levels of interest in the reform efforts were evident. While some faculty members talked of incremental improvement of the existing program, others imagined a more radical restructuring and still others acted on that idea by developing and implementing an alternative teacher preparation program. During a faculty and administrative retreat focusing on reform, we discovered political activity among participants and noticed diverse emotions and levels of interest (e.g., resistance, apathy, curiosity, and confusion) when asked to generate a common vision of the college.

Data Collection

This case study follows Erickson's (1986) interpretive approach and endorses concepts of reality and knowledge consistent with his view. He argues that conceptions of reality cannot be meaningfully separated from the social environment in which they occur. In this sense, qualitative research is holistic and based on the notion of context sensitivity. A basic assumption in interpretive theory is that the formal and informal social systems operate simultaneously. Individuals in everyday life interpret actions in terms of both official and unofficial definitions of status and role. The task for the researcher then, is to try to understand the way participants constitute environments for each other in their interactions and to document the social and cultural organization of the observed events.

To accomplish this task, we conducted the research as participant observers. During the initial phase of the study, we attended and observed restructuring efforts in the College of Education for two years. Previous surveys administered to COE alumni, course syllabi, and other relevant COE documents were also examined. One author held dual roles in the College of Education: a graduate student and the Associate Dean's research assistant. As part of her role, she participated in restructuring meetings, worked with COE administrators to determine an effective evaluation design, and conducted faculty and alumni interviews. The graduate student's dissertation chair was also involved in this study and functioned as a principal investigator.

During the second phase of the study, we identified the following five major constituencies that have interests in the teacher education program in the College of Education: (a) Teacher Preparation (TP) faculty in the COE; (b) faculty in the alternative teacher preparation program (ATP) in the COE; (c) the Department of Education (DOE); (d) the Holmes Group; and (e) principals of schools in which graduates of TP are placed.

It was generally accepted that faculty were responsible for the development and delivery of the curriculum in higher education. The College of Education faculty (TP) therefore comprised the first constituency. Although the TP faculty had constituted the single, dominant teacher preparation program within the college for a number of years, an alternative program broke out for the first time in 1993. Faculty who shared a constructivist perspective and similar beliefs about learning, teaching, and child development (different from the beliefs of the traditional program), began a discussion group to explore alternative approaches to teacher preparation. Enabled by a grant, this group established a pilot alternative teacher preparation track. College faculty from the alternative teacher education program (ATP) comprised the second constituency.

The state Department of Education (DOE) comprised the third constituency. As the primary agent certifying teachers, the DOE purported to influence teacher preparation programs in state universities. In order to have university institutional recommendations for certification recognized by the state, the university must conform to DOE requirements. The DOE identified the proficiencies that a beginning teacher must meet and expected the colleges and universities in the state to develop programs to instill those proficiencies in prospective teachers. We chose the Holmes Group, a voluntary association of teacher preparation programs in research universities, as the fourth constituency in the study. Founded by a group of education deans from such universities, the Holmes Group documents (e.g., *Tomorrow's teachers: A report of the Holmes Groups*, 1986) and guidelines for teacher preparation influenced the thinking and discussion of administration of faculty of the college. Unlike many colleges of education, this COE did not participate in the National Council for Accreditation of Teacher Education (NCATE), the primary accrediting agency. Thus, the Holmes group functions as the principal voice of the profession of teacher preparation. District principals, the fifth constituency, are the primary employers of teacher preparation graduates. In their role as evaluators of new teachers, they were judges of the college's finished product.

Having established the five constituencies, we designed a study, which became phase three. Following Erickson's (1986) advice, we employed multiple methods and generated rich descriptions in order to provide a credible, internally consistent account of constituent ideologies. We conducted formal and informal interviews, took detailed notes of participant interactions and activities, and examined archival data and artifacts.

Interviews

We conducted 23 semi-structured interviews during a one-year period with representatives from the five constituent groups. The interviews were distributed as follows: five interviews each with the TP, ATP, Department of Education, and district principals and three interviews from the Holmes group. We chose a semi-structured format, the intent of which was to reveal the multiple perspectives of members of diverse constituent groups. Through the course of the interviews, we raised questions that would reveal the informants' images of schooling, curriculum development, teacher training, and instructional practice.

The following are representative of questions we developed for individual and group interviews:

1. Describe the type of classroom you believe graduates of teacher preparation programs should be able to organize.
2. Describe the type of classroom management philosophy you believe graduates of teacher preparation programs should be able to implement.
3. Describe the type of classroom materials you believe graduates of teacher preparation programs should be able to select and use.
4. Describe the type of knowledge you believe graduates should have about the community served by the school.
5. Describe the kind of knowledge you believe graduates should have about working with pupils from different language and ethnic groups.
6. Describe the kind of knowledge you believe graduates of teacher preparation programs should have about the socio-political nature of teaching.
7. Describe the type of ideology you believe graduates of teacher preparation programs should have about teaching and learning.

Although the list framed the questions, we avoided interfering or directing participant answers. Because we allowed interviewees to discuss issues off the list, we relied on a discovery-oriented, inductive approach to interviewing (Bernard, 1994).

In addition to individual interviews, we participated in four different focus groups conducted as a part of the Dean's reform initiative. The institution's graduates now teaching, district principals, and faculty from the traditional and alternative teacher preparation programs attended the group interviews. Transcripts from those interviews formed an alternative data source to the individual interviews. The focus group involved the systematic questioning of several individuals in a formal setting (Drever, 1995; Merton, Fiske, & Kendall, 1956).

Documents
Although the interviews were the heart of the data collection, we also collected documents and archival records to provide an alternative perspective on the research question. Marshall and Rossman (1989) argue that the unobtrusive nature of document and archival record collection provide a rich data source without disrupting the site. Included in the data were such documents as the state's Department of Education's Language Arts Essential Skills," the Holmes group publications, and the COE teacher preparation course syllabi.

Observational Data
One of the authors attended all COE faculty restructuring meetings; prepared faculty fora on teacher preparation; and conducted focus groups with school principals, teacher preparation program alumni, faculty, and cooperating and supervising teachers. During this phase of data collection we gathered important background knowledge for the study that would have otherwise been considered confidential and unavailable for student use. Although we chose not to tape every meeting, these "brainstorming" sessions with faculty members provided additional insight for the phenomena under study.

Analysis of Data

We began the study with the idea that the educational system comprises many constituent groups and the micropolitical hypothesis that diverse constituencies have different ideologies regarding schooling (e.g., images of curriculum, teachers and teaching, pupils, and teacher preparation). We collected data that might reveal images of schooling and test the research hypothesis.

Because of the emphasis placed on induction and intuition, we allowed meanings and definitions to emerge. All final categories and assertions were grounded in interview transcripts, observations, and document data. This method allowed for some discovery in data generation and analysis and helped guard against confirmatory bias. Notes were worked over after all observations and interviews were completed, thus allowing for a full picture of what occurred and providing a greater opportunity to encounter disconfirming evidence.

Padilla's (1991) concept modeling methodology was used as a strategy for organizing and displaying the data. According to Padilla, one way to explain a situation is to identify various assumptions contained in the data and organize them into a coherent whole. In the concept modeling method, assertions contained in the data were fundamental elements for analysis. First, we created a matrix in which to arrange the concepts, namely, images of curriculum, teachers, pupils, and the like. Next, we reduced long statements from interview transcripts and excerpts from documents to short paraphrases, and entered these data into appropriate cells of the matrix. After we observed how data were arrayed across the constituent groups, we highlighted areas of convergence and divergence among constituent images.

Images and Evidence

On the basis of the concept model we find that fundamental differences exist in the five constituencies in the images of the curriculum (i.e., purpose, origin, organization, and content), teacher, pupils, and teacher education. Table 1 and Table 2 show the concept matrix with paraphrases inductively derived from data excerpts, illustrative samples of which are then presented and interpreted. Table 1 and Table 2 are organized as a matrix of five constituencies by the four topics.

Images of Curriculum

The data from this study suggest divisions in regard to curriculum content, purpose, and subject areas for curriculum development. Among constituencies, image of the curriculum differed along four dimensions: purpose, origin, organization, and content. Furthermore, the purpose of curriculum affected the other three dimensions. That is, notions of purpose corresponded with notions of appropriate curriculum content and how and where curriculum should be developed.

Table 1
Images of Curriculum

	Teacher Preparation (TP)	Department of Education (DOE)	Principals	Holmes	Alternative Teacher Preparation (ATP)
K-12 Curriculum Purpose	*Competency* Essential skills define what pupils ought to know. The districts use those skills for curriculum alignment.	*Competency* State Essential Skills dictate the standardized skills and competencies all students should learn.	*Competency* The curriculum is determined in every district and is a statement of goals. Every district really has the same goals.	*Inquiry* Curriculum should emphasize collaborative learning, reflection, and dissemination of new and changing knowledge about teaching & learning (Holmes, 1992).	*Inquiry* Students explore critical, reflective thinking and engage in the examination of what is considered valuable knowledge in schools. Students study curriculum from many perspectives.
Origin	*Centralized* Without curriculum guidelines it is too easy for novice teachers to make independent decisions about what they prefer to teach.	*Centralized* Curriculum standards are viewed as good.	*Centralized* Standardized curriculum is preferred, one that results in a collection of measurable skills; curriculum is all prescribed.	*Context dependent* Real knowledge is purpose-built, site-built and infused with the learners' sense of purpose. Knowledge is imaginatively constructed, not passively acquired (Holmes, 1990).	*Context dependent* Teachers should negotiate with the student a set of activities relevant and meaningful but also challenging. They design a curriculum rather than latching on to a standardized one.
Organization	*Molecular* Content curriculum is dealt with in departments, with strict divisions among content areas.	*Molecular* Essential Skills specify content and skills for each grade level. General skill areas are within the disciplines	*Molecular* There is a need for practical learning, instead of the holistic orientation not implemented in the public schools.	*Integrated* The basics are not just facts but also concepts and relationships. Concepts and facts merely make up a related background and foreground (Holmes, 1990).	*Integrated* Curriculum is holistic where all learning emanates from pupils' interests.
Content	*Technical* Curriculum must be politically and religiously neutral. Values can be eliminated from the curriculum. Values should not be taught in school.	*Technical* Values should not be taught in school. They should be taught in the church.	*Technical* Pupils have to determine what's right and wrong. The concepts should be taught in the home, not by the teachers.	*Social* In transmitting knowledge, you give students more than math and science. You're transferring a whole value system that is ingrained in the literacy system (Holmes, 1990).	*Social* Included in the content is the hidden curriculum and a social and political curriculum. Political and ethical agendas are inherent in school instruction as evidenced by what is included or excluded.

We identified two recurrent, diverse perceptions regarding the purpose of curriculum: "competency based" and "inquiry based." Those who held a competency-based perspective felt pupils should learn a predetermined set of content-independent cognitive skills applicable in a variety of situations. Advocates primarily focus on refining intellectual operations or understanding the processes by which learning occurs in the classroom. The competency based conception of curriculum highlights the intellectual processes rather than educative context and content. In contrast, those with an "inquiry based" perspective believed instruction should force pupils to critically analyze what they learn. Additionally, proponents believe that instructional content and the learning context are interdependent. They often define curriculum in process terms such as creative or critical thinking, metacognition, and experimentation (ATP interview).

We discovered further dichotomies within the other three dimensions (i.e., origin, organization, and content): 1) standardized versus local curriculum development; 2) fragmentation versus integration of subject areas; and 3) exclusion versus inclusion of political issues. However, curriculum purpose was a common thread, if not a dominant theme, among other dimensions. Those with a competency-based image saw curriculum as officially constructed (origin), organizationally fragmented (organization), and politically neutral (content).

Standardized versus local curriculum development. Advocates of the "competency" stance assumed that the purpose of curriculum was to teach students a clearly defined set of competencies dictated and determined by official governmental standards. Constituents viewed curriculum as a hierarchical set of basic skills, which students must learn. Advocates assume district officials find, define, and dictate this knowledge to practitioners, and that the prescribed facts and guidelines constitute the best curriculum model.

Conversely, those who held an "inquiry based" stance questioned the existence of a unique body of knowledge and challenged the assumption that individuals external to the classroom possess more relevant expertise than teachers. They believed that teachers were capable of identifying knowledge and, more significantly, that knowledge was inherent in the learning context. Thus, curriculum ceased to be static, predefined set of skills and outcomes, but rather functioned as a dynamic, evolving system unique to each classroom environment.

Fragmentation versus integration. The fluid nature of an evolving system required a cohesive organizational approach to curriculum knowledge. Constituent groups recognized contrasting strategies to arrange and teach school content. The molecular-holistic dichotomy, which represents the dilemma between the fragmentation and integration of content, surfaced in all data sources. A molecular conception reflects an assumption of strong classification among content areas and the holistic conception assumes weak boundaries (Ginsburg, 1986).

Exclusion versus inclusion of social issues. Another recurring dichotomy was the "technical" versus "social" dilemma, which represented contrasting preferences in curriculum content. Those who espoused the "technical" view avoided what they felt were value-laden issues opting instead to assume a neutral position devoid of specific social issues. Constituents who embraced the "social" perspective assumed a connection between the order of society and what the curriculum of schools in that order contained.

ATP faculty argued that depicting schooling and curriculum as neutral and apolitical systems masks the bias in content selection and provides a facade of objectivity and fairness. To foster a balanced perspective, curricular content and classroom experiences should include political, social, and ethical issues. Moreover, they felt these constructs were inherent in schooling because, at some point, someone determined which facts were relevant.

Images of Teachers

What social expectations should apply to those who hold teaching positions? Most concede that teaching is a complex, multifaceted act that requires numerous activities, behaviors, and decision-making abilities (Barnes, 1989; Spring, 1985). Although teaching requires knowledge of content and the ability to apply that knowledge in diverse settings there is room for differences in emphasis in this general description, even contradiction and controversy.

The five constituencies held alternative images for what teachers do and what they should know. Moreover, we found a series of metaphors for what constitutes teachers' work. These emerging metaphors were so robust that we chose to use them as a device to organize, analyze, and present the data.

We uncovered two dominant, conflicting images of teachers, specifically "teacher as technician" and "teacher as inquirer." Constituents advocating a competency-based conception of curriculum saw teachers as "technicians" who assume all responsibility for classroom learning and management. Therefore, teachers, serving as diagnosticians and transmitters of knowledge, manipulate forces to produce predictable ends (Combs, 1991).

In contrast, proponents of an "inquiry based" curriculum view teachers as facilitators in the learning environment. Curriculum goals are broad and context-dependent, and all participants in the classroom share the responsibility for student outcomes. That is, teachers and pupils collectively engage in decision-making processes, and as a result, the shared responsibility relieves the teacher's management burden. Rather than functioning as managers, teachers function as consultants in the students' evolving learning processes.

Table 2
Images of Teachers, Pupils, and Teacher Preparation

	Teacher Preparation (TP)	Department of Education (DOE)	Principals	Holmes	Alternative Teacher Preparation (ATP)
Teachers	*Technicians* Teachers should have a backpack of methods to deliver material. Content is presented to meet the needs of diverse students. Evaluator - Teachers effectively evaluate learners.	*Technicians* Experts - Teachers should be multifaceted but certainly they must be experts in their content area. Evaluator – Teachers evaluate what pupils have done to assess mastery.	*Transmitters* Diagnostician - Teachers need to diagnose the correct level of difficulty each student needs in relation to curriculum standards.	*Inquirers* Teaching should include inquiry, reflection, documentation, and dissemination of new and changing knowledge (Holmes, 1992).	*Inquirers* Teachers are inquirers. They must question what's going on in the lives of students and society. Facilitator - Teachers act like coaches in the learning environment. Their role is to enable

					learners to construct their own knowledge.
Pupils	*Deficient* Pupils are empty bank accounts when they come to class.	*Deficient* Pupils fall short because of the baggage they carry with them. At-risk students are those that have extra baggage, whether it's their parents or the community.	*Deficient* Skills are arranged hierarchically so that higher order thinking or problem solving is pursued once basic skills are mastered.	*Resource Centers* Learning is an active process in which children construct and reconstruct knowledge. Knowledge is imaginatively constructed, not passively acquired (Holmes, 1990).	*Resource Centers* Pupils bring a variety of experiences and abilities with them, which must be considered in curriculum design. Pupils must think critically and differently from others.
Teacher Preparation	*Competency* The model is based on expanding an individual repertoire of well-defined, classroom practices. After graduates are armed with the best foundation we send them into the field.	*Competency* District and governmental guidelines are set up which determine what university students should experience in teacher education.	*Competency* Teaching must conform to district curriculum competencies and standards. Preparation programs should diagnose district patterns and translate them into the curriculum.	*Inquiry* Investigation, critical reflection, and inquiry are central features of teacher education (Holmes, 1990).	*Inquiry* The purpose of teacher education is to prepare teachers to be inquirers in the classroom. Teachers should challenge student beliefs and ideologies to make them critical thinkers.

Teacher as technician. Principals, DOE officials, and TP perceive teachers primarily as expert technicians, who diagnose the learning situation, select techniques needed to reach goals, transmit content by sequencing and fragmenting chunks of information, and evaluate the outcome to determine if objectives were achieved. This image assumes: 1) a top-down organization in which teachers teach a set of competencies dictated and determined by official standards; 2) teachers possess content knowledge and are solely responsible for transmitting knowledge to students; and 3) learning is the development of competence as evidenced by learning standardized skills. Teachers, rarely in the business of creating new knowledge and others, such as universities and scientists, for example, are knowledge producers.

Teacher as inquirers. ATP and the Holmes Group viewed teachers as "inquirers" assisting and consulting students in an on-going process of exploration and discovery and placing more emphasis on participation and joint responsibility in the learning process. In this capacity teachers promote inquiry and function as facilitators in the classroom. In addition, a teacher's expertise lies in the promotion of practices that create the conditions for social change. Constituents who view "teachers as inquirers" assume teachers are active in their pursuit of professional growth and reform and in their construction of and orientation to curriculum (ATP interview). This image assumes a bottom up organization in which curriculum is developed by the teacher to fit the needs of students in a specific context. Advocates of this image assume teachers are autonomous and are professionals who need to exercise more influence over their work rather than conforming to arbitrarily assigned tasks.

Images of Pupils

Images of pupils were compatible with constituent images of teachers. For example, those who viewed "teachers as technicians" assumed pupils were passive recipients and dependent learners in the classroom. Practitioners also assume that students must first master a set of competencies before they attempt critical thinking. Therefore, students rarely engage in decision-making processes or employ discovery methods. In contrast, constituents who viewed "teachers as inquirers" thereby allowed pupils to assume more responsibility for creating their own knowledge. Because teachers and pupils are involved in explorations, more solutions are possible. The ATP learning context emphasizes problem solving and abstract thinking rather than prescribed solutions. Teachers incorporate student diversity (e.g., gender, ethnic, language, and academic) into the curriculum, which is thematic and negotiated. Thus, preservice teachers are more able to serve a culturally diverse population.

Images of Teacher Preparation

Is there one best way of preparing teachers? Should all preservice teachers receive the same body of knowledge and have the same experiences? We attempted to answer these questions by interviewing individuals involved in the day-to-day operation of schools and programs of teacher preparation within the study site.

Images of teacher preparation were distilled across constituent views of curriculum, teachers, pupils, and schools. The first image reflected the belief that official, governmental experts determine what skills and abilities a new teacher should possess. These competencies, which are consistent with established policies and the state curriculum developed

BEST COPY AVAILABLE

by DOE administrators, should be learned in university classroom settings and field experiences. Teacher preparation for competency's framework is found in behavioral psychology and practitioners are expected to control stimuli to produce predetermined outcomes (Combs, 1991). Subsequently, responsibility for direction is in the hands of teachers, which dictates a passive, conditioned role for the learner.

Conversely, constituents with an inquiry-based perspective saw teacher preparation as a process-oriented framework to encourage critical thinking, responsibility, and responsiveness toward a diverse student population using broadly defined subject matter and goals. Therefore, faculty concentrate on creating optimal conditions suitable for exploring (Combs, 1991).

We found a list of competing interests, needs, and ideologies among Teacher Preparation Program constituents. Many of these discrepant views coalesce around the issue of what we assess as relevant and what balances we strike among academic knowledge, technical competence, and critical inquiry. Whereas advocates of technical competence have always prided themselves on its direct relevance to the workplace, others have valued a broader education including the exploration of disciplinary knowledge and the development of higher order skills such as critical thinking, problem solving, communication, and research skills (Clancy and Ballard, 1995). Yet, universities and teacher preparation programs remain under pressure to be more accountable to the workplace and to principals who hire graduates.

To make a more concise argument we chose to epitomize basic competing ideologies in two idealized, hypothetical classrooms. In Classroom A, which is representative of TP, COE, and principals' perceptions, teacher preparation instructors use a knowledge-based curriculum for prospective teachers. The syllabus delineates course objectives, goals, and assignments and grades are determined by content competency predetermined by the instructor. Students are taught from a behaviorist approach, which assumes knowledge is fixed, technical, and generated from an external source. In Classroom A, teachers, as consumers and transmitters of knowledge, are primarily responsible for student learning.

In Classroom B, which represents The Holmes Group and ATP ideologies, teacher preparation students make decisions about course design and content, identify an independent ethnographic research project relevant to their needs, and study the whole language method. Using a thematic approach, they design lesson plans that may include history, literature, and art. The instructor argues that a standardized curriculum is ineffective because of the unique nature of each teaching and learning context. From this perspective, teachers function as constructors and facilitators of knowledge, while pupils assume more responsibility for their learning (i.e., are also inquirers). Teachers emphasize the constructivist assumption that learning is social behavior; therefore, classroom activities must include interactive processes that promote broad conceptual understanding. Teachers value knowledge as problematic, holistic, negotiated, and socially relevant. This constituency highlighted the interpretive value of experiential knowing.

The existence of these two diverse perspectives within a common teacher education program highlights intra-organizational activity and supports micropolitical theory that assumes that any complex organization, such as an institution of higher education, comprises several constituencies that contend with each other over resources, interests, and definitions of schooling. As we examined this teacher preparation program, we saw one coalition develop (ATP), due to common ideologies, and break away from the traditional program (TP). Although faculty ideologies within the ATP were unitary, their educational vision was different from the traditional program. In fact, ATP faculty believed that their vision was "Not for everyone" and did not intend to market their program as the model for systemic reform in the College.

While we interviewed and observed TP faculty, we uncovered a main voice, which reflected common ideologies within that constituency. We later presented this dominant voice in the matrices and text. However, there were additional, divergent voices within that constituency. For example, some TP faculty shared common views with the ATP program; however, they chose not to break away. Other responses were considered rare and were perhaps anomalies within the TP. Based upon further splintering of ideologies, we could have divided the TP into several other subcategories.

Conclusions

Systemic reform aims to restructure the entire educational system, a process requiring the organization of all facets of schooling. To facilitate dynamic educational change, the process must involve all intersecting components of the system. However, if one of these components contradicts the others, change in the remainder of the system is subverted or distorted. According to systemic reform, solutions must come through the development of shared and negotiated meanings (Fullan, 1991) as the Dean of the College of Education had desired in attempting to institute his unitary vision of teacher preparation reform.

Edelman (1995) argued that reform that enforces a central, standard vision of an organization is likely to have only symbolic, rather than instrumental effects if one assumes that individuals act (e.g., teachers) toward objects (e.g., courses) according to the meanings and definitions of the situation that those objects have for them (Blumer, 1969). Then the imposition of a unified vision or central reform will only provoke resistance, increase teachers' political action, drive ideological differences underground (Benveniste, 1989), or result in a splintering of subprograms that are internally consistent but contradictory across subprograms.

311

The present study of a single case, the efforts of a COE to reform its teacher preparation program, revealed several of the organizational processes above and highlighted the fact that systemic reform is unobtainable. Through noncompliance, instructors will openly resist or dramatically revise policy with which they are ideologically opposed (Blase, 1997; White & Wehlage, 1995) and reform will be "subverted by the complex interplay of human transactions that do not happen to fit the printed scenario" (Benveniste, 1989, p. 329).

Reform can disrupt the status quo because the increased expectations make schools and their communities unavoidably more political. Political actors (e.g., teachers) rush in to take advantage of openings, grabbing control of agendas and resources in the temporary vacuum created by reform. When individuals form groups, they validate their ideologies and strengthen their position. As a consequence of reform, teachers may realize increased political power.

An intra-organizational process in the present case study linked this micropolitical assumption with empirical evidence. Although the development of a common vision within the college or systemic reform was not achieved, one coalition successfully implemented a reformed teacher preparation program based upon their shared images of schooling. The ATP group, which separated from the traditional TP track, enjoyed increased political power and developed a teacher education program that reflected components of the dean's restructuring vision and the national reform agenda. For example, ATP increased active learning environments, promoted collaboration between the university and local schools, and educated preservice teachers to better serve the needs of a culturally and linguistically diverse student population. The Holmes Group (1992), the Carnegie Task Force, and the National Commission for Excellence have underscored the importance of educating teachers to understand and serve the needs of a diverse population and have identified this as an important goal in teacher education reform.

To what other institutions can one generalize these findings? Their consistency with other research suggests the case is not unique (Ball, 1987; Blase, 1991, 1997; Blase and Anderson, 1995; Hoyle, 1999; Lindle, 1999; Malen, 1994; Mawhinney, 1999; Noblit et al., 1991; Scribner et al., 1995; and West, 1999). Yet, what is generalized from case studies is not the empirical features of this particular college of education. Instead, as Ragin and Becker (1992) pointed out, what is generalized is the theoretical process discovered here. That is, the existence of constituencies within institutions and the conflicting and discrepant ideologies across constituencies create program and policy incoherence, and this intra-organizational process may occur in other institutions as well.

From a micropolitical perspective, is change possible? Change occurs in organizations because of internal processes, practices, and conflict. Conflict and contrasting views of schooling in this case prompted and enabled change. Reform in one track of teacher education (ATP) was an expected by-product in this social system comprising competing ideologies. Combs (1991) pointed out that organizational reform stems from changes in the beliefs of the people at the street level (e.g., teachers) and because educational reform concerns individuals in a culture we must create a system specifically designed for the "human problem" (p. 148). He suggested a move from a closed to an open system. In the open system, for teachers, the basic shift entails movement toward a student-centered view of learning (Levin, 1994) while administrators function as facilitators rather than managers.

Recognizing and addressing the human problem in reform requires changes not only in the structure and administration of, but also in how we perceive the organization. Ball (1987) added that the focus on organizational matters should be augmented by a parallel focus on the content of policy and decision-making in schools since a large portion of the content is ideological. Even when goals are clearly delineated, different educational and political ideologies may lead educators to approach their tasks from diverse directions. As Ball pointed out, it is "possible to find enormous differences between subject departments within the same school and even between teachers in the same department." (p department." (p 14)

The increased expectations and political activity associated with education reform make the study of micropolitics absolutely crucial for school administrators and reform advocates (Lindle, 1999; Mawhinney, 1999; West, 1999). The empirical evidence presented in this case, which is consistent with prior work in the field, supports the assertion that all organizations are composed of coalitions and individuals with competing ideologies; as Barr-Greenfield (1975, p. 65) so poignantly stated, that "is the organization." In other words, reformers have to acknowledge and work with the limitations and opportunities inherent in the processes of such a structure. Micropolitical advocates such as Bacharach (1996) argued that current theories of organizational change fail to pay adequate attention to how organizations move from one stable state to another. Therefore, a model of the organizational transformation process or intra-organizational practices must also be examined for those considering reform in education.

References

Bacharach, S. B. (1996.). The organizational transformation process: The micropolitics of dissonance reduction and the alignment of logics of action. *Administrative Science Quarterly, 41*(3), 477-506.

Ball, S. J. (1987). *The micropolitics of the school: Towards a theory of school organization.* London: Methuen.

Barnes, A. (1989). Structuring knowledge for beginning teachers, in M. Reynolds (ed.), *Knowledge base for the beginning teacher*. Oxford, England: Pergamon, 13-22.

Barr-Greenfield, T. (1975). Theory about organization: A new perspective and its implications for schools, in V.P. Houghton, G.A.R. McHugh, & C. Morgan (ed.), *Management in education* London: Wardlock/Open University Press, 69-87.

Benveniste, G. (1989). The micro-politics of the school: Towards a theory of school organization (Review). *Educational Administration Quarterly*, 2, 326-328.

Bernard, H. R. (1994). Unstructured and semistructured interviewing. *Research methods in anthropology: Qualitative and quantitative approaches.* Thousand Oaks, CA: SAGE.

Blase, J. (1991). *Politics of life in schools: Power, conflict, and cooperation.* London: Newbury Park.

Blase, J. (1997). The micropolitical orientation of facilitative school principals and its effects on teachers' sense of empowerment. *Journal of Educational Administration*, 35(2), 138-164.

Blase, J. & Anderson G. (1995). *The micropolitics of educational leadership: From Control to Empowerment.* New York: NY: Teachers College Press, Columbia University.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method.* (Englewood Cliffs, NJ: Prentice Hall).

Clancy, J. and Ballard, B. (1995). Generic skills in the context of higher education. *Higher Education Research and Development*, 14(2).

Combs, A. W. (1991). *The schools we need: New assumptions for educational reform.* Lanham, MD: University Press of America, Inc.

Drever, E. (1995). *Using semi-structured interviews in small-scale research. A teacher's guide.* Scottish council for research in education, Edinburgh.

Edelman, M. (1985). *The symbolic uses of politics.* Chicago, IL: University of Illinois Press.

Erickson, F.(1986). Qualitative methods in research on teaching. In M. Wittrock (ed.), *Handbook of Research on Teaching.* New York: Macmillin Co, 114-16.

Ginsburg, M. B. (1986). Reproductions, contradictions, and conceptions of curriculum in preservice teacher education. *Curriculum Inquiry*, 16(3), 283-309.

The Holmes Group.(1986). *Tomorrow's teachers: A report of the Holmes Group, Inc.* East Lansing, MI: The Holmes Group, Inc.

The Holmes Group,(1990). *Tomorrow's schools.* East Lansing, MI: The Holmes Group, Inc.

The Holmes Group, (1992). *Recommitting and reshaping the education school: Stimulus paper.* East Lansing, MI: The Holmes Group, Inc.

Hoyle, E. (1999). The two faces of micropolitics. *School Leadership and Management*, 19(2), 213-222.

Levin, B. (1994). Improving educational productivity: Putting students at the center. *Phi Delta Kappan*, 75(10), 758-760.

Lindle, J. C. (1999). What can the study of micropolitics contribute to the practice of leadership in reforming schools? *School Leadership and Management*, 19(2), 171-178.

Mawhinney, H. B. (1999). Reappraisal: The problems and prospects of studying the micropolitics of leadership in reforming schools. *School Leadership and Management*, 19(2), 159-170.

Malen, B. (1994). The micropolitics of education: Mapping the multiple dimensions of power relations in school politics. *Journal of Education Policy*, 9(5), 147-167.

Marshall, C., & Rossman, G. B. (1989). *Designing qualitative research.* Newbury Park, CA: SAGE Publications, Inc.

Merton, R. K., Fiske, M., & Kendall, P. L. (1956). *The focused interview.* Glencoe, IL: Free Press.

Noblit, G., Berry, B., & Dempsey, V. (1991). Political responses to reform: A case study. Education and Urban society, 23(4), 379-395.

O'Day, J. A. & Smith, M. S. (1993). Systemic school reform and educational opportunity. In S. Fuhrman (ed.), *Designing coherent education policy: Improving the system*. San Francisco, CA: Jossey-Bass Publishers, 250-311.

Padilla, R. V. (1991). Using computers to develop concept models of social situations. *Qualitative Sociology, 14*(3), 263-274.

Ragin, C. C. (1992). "Casing" and the process of social inquiry. In C. C. Ragin & H. S. Becker (Eds.), *What is a case?* Cambridge, MA: Cambridge University Press, 217-226.

Ragin, C. C., & Becker, H. S. (1992). *What is a case?* (Cambridge, MA: Cambridge University Press).

Salamon, G. & Thompson, K. (1973). *People and Organizations*. London: Longman.

Scribner, J. & Layton, D. (1995). The study of educational politics. 1994 Commemorative yearbook of the politics of education association (1969-1994). Bristol, PA: Falmer Press.

Spring, J. (1985). *American education: An introduction to social and political aspects*. White Plains, NY: Longman.

West, M. (1999). Micropolitics, leadership and all that. The need to increase the micropolitical awareness and skills of school leaders. *School Leadership and Management, 19*(2), 189-195.

White, J. A., & Wehlage, G. (1995). Community collaboration: If it is such a good idea, why is it so hard to do? *Educational Evaluation and Policy Analysis, 17*(1), 23-38.

About the Authors

Susan Haag
College of Engineering and Applied Sciences
Arizona State University
Tempe, AZ 85287
Email: susan.haag@asu.edu

Susan Haag is the Director of Assessment and Evaluation for the College of Engineering and Applied Sciences at Arizona State University. She advises and assists faculty in the development and implementation of course and program assessment and evaluation. Her academic and personal research focuses on educational reform, program evaluation, integrating assessment and technology into the curriculum, organizational policy, and recruiting and retaining underrepresented populations.

Mary Lee Smith
College of Education
Arizona State University
Tempe, AZ 85287

Email: mlsmith@asu.edu

Mary Lee Smith is Professor of Educational Policy Studies in the College of Education, Arizona State University and also Professor of Methodological Studies. In her early research career she worked on methodology of meta-analysis, particularly the meta-analysis of research on psychotherapy effectiveness. She has spent a number of years working on alternative methodologies in evaluation and policy research and has applied them to study assessment policies and policies to end social promotion.

315

Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Affirmative Action in Higher Education:
An Analysis of Practices and Policies

Alfred R. Cade Jr.
Missouri Southern State College

Abstract

This study analyzed the variations of policies and practices of university personnel in their use of affirmative action programs for African American students. In this study, the policy topic is affirmative action and the practices used in admissions, financial aid, and special support services for African-American students. Surveys were mailed to 231 subjects representing thirty-two Missouri colleges and universities. Most of the survey respondents were male, white, and nearly two-thirds were above the age of forty. Ethnic minorities were underrepresented among the professionals. Seventy-two percent of respondents were white, 23% were African American, and 5% were Hispanic. The results of this study suggest a positive picture of student affirmative action practices and policies used by Missouri personnel. Differences among professionals were at a minimum. The overall mean score for support in diversifying Missouri institutions was fairly high, and this may reflect diversity initiatives taken by the Missouri Coordinating Board for Higher Education in the late 1980s, and early 1990s. Data suggested that Missouri personnel are aware of the judicial scrutiny by the courts in administering student affirmative action. Most Missouri institutions use a single process for assessing all applicants for admission, without reliance on a quota system. The recent Hopwood decision showed little impact on the decisions regarding professionals' use of student affirmative action at Missouri institutions. Although public attitudes toward student affirmative action may play a role in establishing policies and practices, Missouri personnel are very similar in their perceptions regardless of race/ethnicity, gender, and institutional office or position.

Introduction

The purpose of analyzing race-based affirmative action practices used by higher education personnel was based on concurrent court rulings and the political climate. California, Washington state, and Florida ceased the use of affirmative action practices in higher education. In the court decision, *Hopwood v. State of Texas*, (1996), the court rendered their decision that ended race-based affirmative action practices historically used by colleges and universities in the Fifth District. Some speculate that these actions precipitated reactions by institutions of higher education in their approach to practices and policies concerning affirmative action.

According to Cross and Slater (1997), analyzing the use of affirmative action practices and policies regarding minority access to higher education is important for the future of our country. Both authors' calculations suggested that if standardized tests become the single norm in admission decisions, African-American enrollment at some institutions will drop by at least one half and in some cases as much as 80 percent.

Former higher education administrators Bok and Bowen (1998), concluded in their longitudinal study that race-neutral standards would produce troubling results in the proportion of African American students in higher education. Statistics at the University of Texas at Austin, School of Law indicated a decrease in the number of applications from African-American students following the Fifth Circuit Court's decision in Hopwood (Henry, 1998, Cross & Slater, 1997, Chenoweth, 1997).

In the University of California System following the passage of Proposition 209 (the California Civil Rights Initiative), African-American applications and admission declined significantly (Jones, 1998). In the spring of 1998, the U. S. House of Representatives voted 249 to 171 to reject an amendment, which if passed could have barred federal support for public colleges and universities that granted preferential treatment in admissions based on an applicant's race, gender and ethnicity (Burd, 1998). Consequently, universities are evaluating their affirmative action policies and practices used in student admission and retention. For these institutions, lawsuits and political ramifications forced some to defend and to abandon the use of race in their policies (Kurlaender & Orfield, 1999). Are colleges and universities altering their practices and policies in using race as a criterion in admissions, financial aid, and special support services? Could this disparity widen if colleges and universities altered their practices and policies in the use of affirmative action?

Various public opinion surveys consistently found that most Americans valued and embraced diversity whether in the workplace, or university setting. Americans are more inclined to modify than dissolve existing race-based policies (Bolden, Goldberg, & Parker, 1999). Universities inevitably understood that having a diverse student body was essential for student growth. This cultural and ethnic educational environment has naturally effected the outcomes of learners in a university setting. In regards to race conscience efforts, decision makers in higher education are left pondering over decisions on what ways to promote inclusion and diversity.

Purpose

The purpose of this study was to analyze the variations of policies and practices of selected Missouri college and university personnel in their use of affirmative action programs for African American students. In this study, the policy topic is affirmative action and the practices used in admission, financial aid, and special support services for African-American students. At the time of this study the courts have not mandated Missouri institutions to alter their admissions and financial aid policies in affirmative action procedures.

This study analyzed the present use of affirmative action policies and practices being administered for student admission, financial aid, and special support services by selected colleges and university personnel in Missouri. Affirmative action policies are currently being challenged at a vast number of colleges and universities across the nation. Institutions of higher education are concerned with the strict scrutiny of the courts in reference to practiced affirmative action policies (Kurlaender & Orfield, 1999).

Over the past few years, numerous books, articles and scholarly journals addressed the issue of affirmative action, mainly concerning college admissions and financial aid. Nearly all these reports dealt with the legal, ethical, and political issues surrounding affirmative action and preferential admissions for students of color (Bolden, Goldberg, & Parker, 1999). Very few of the studies attempted to forecast how the attacks on affirmative action influenced the policies and practices of those in academia (Bowen & Bok, 1998).

In essence, the present study was significant given the fact that institutions should consider the condition for African Americans students in higher education if we began to eliminate institutional affirmative action policies and procedures. In the late 1980s, the Missouri Coordinating Board for Higher Education (CBHE) developed strategies to increase minority recruitment and retention in Missouri institutions of higher education. Their report entitled "Challenges and Opportunities: Minorities in Higher Education" urged Missouri institutions to develop policies and practices to address the issue of low minority participation (Missouri CBHE Review, 1988). In general, African American students are more likely than white students to come from educational backgrounds that will not adequately prepare them for the challenges of post secondary education (Bowen & Bok, 1998). The objective of the (CBHE) report was to have an impact on the goal of diversifying Missouri society, particularly in the middle and upper reaches of the socioeconomic status system.

The CBHE report, and other prominent educational publications, need to analyze affirmative action policies within institutions of higher education. Have institutions developed policies and practices to address the issue of minority, and in particular, African American student participation? If so, to what degree are personnel using affirmative action practices? Do significant differences exist regarding affirmative action practices used by higher education personnel? These were questions the researcher asked investigated throughout this study.

Hypotheses   **318**

In this study, null hypotheses were developed based on the theoretical support that existed in the literature:

- Hypothesis one: There are no significant differences regarding participants perception toward affirmative action practices among Missouri personnel based upon their institutional affiliation (public or private).
- Hypothesis two: There are no significant differences regarding perception toward affirmative action practices between participants grouped by ethnicity.
- Hypothesis three: There are no significant differences regarding perceptions toward affirmative action practices between participants grouped by gender.
- Hypothesis four: There are no significant differences regarding participants perceptions toward affirmative action practices between Missouri institutions based on admission classification.
- Hypothesis five: There are no significant differences regarding participants perceptions toward affirmative action practices between Missouri institutions based on size of institution.
- Hypothesis six: There are no significant differences regarding perception toward affirmative action practices between participants based on number of years in position at institution.
- Hypothesis seven: There are no significant differences regarding perception toward affirmative action practices between participants based upon position within the institution.
- Hypothesis eight: There are no significant differences regarding perception toward affirmative action practices currently used by Missouri personnel when institutions are grouped according to the percentage of African-American student enrollment.

Method

This study followed a quantitative descriptive approach to investigate the level of variability in affirmative action practices by Missouri institutions. According to Gay and Airasian (2000), quantitative descriptive studies are conducted to acquire knowledge about preferences, practices, concerns, or interests of a specific group. A quantitative descriptive survey was used to collect data on both practices and policies used by the selected population. Data were coded and analyzed to yield the variance that existed among Missouri college and university personnel in their practices of student affirmative action. Following the collection of data the major statistical analysis used was an analysis of variance. The mean scores for the subjects were analyzed to measure the degree of difference that existed among group characteristics.

Following the collection of data the major statistical analysis used was an analysis of variance (One-Way ANOVA). Additionally, the researcher in this study analyzed selected hypothesis using a t-test. Hypothesis 2, 4, 5, 6, 7, and 8 were tested using the One-Way ANOVA. The reason for this was to analyze multiple groups (variables) for comparison. Hypothesis 1 and 3 were analyzed using the t-test for independent samples. Hypothesis 1 and 3 compared two distinct groups. Tukey's HSD post hoc test was used by the researcher to determine which groups significantly differed in their perceptions toward the use of student affirmative action practices and policies. The mean scores for the selected samples were grouped and compared to measure the degree of variance. The .05 level of significance was used for all statistical analysis.

Subjects

The subjects in this study represented Chief Executive Officers (chancellors, presidents, and vice-presidents), Enrollment/Admissions Representatives, Financial Aid Counselors, and Special Support Services staff. The researcher identified approximately 275 subjects representing thirty-two Missouri colleges and universities. Following the examination of institutional flow charts, this number was subsequently reduced to 231. This reduction was based on assessing only professionals having an impact on student and institutional policies and practices. The information obtained from the selected sample pertained to student affirmative action procedures along with demographic information.

Variables

The dependent variables in this study were the selected levels of affirmative action policies and practices used by the subjects in six areas of practices and policies. The independent variables included the sample's demographic information provided in the survey. These independent variables were the subject's position, subject's department or office, type of institution (public or private), admission status, size of institution, and student composition. Additional independent variables were explored related to the subject's gender, ethnicity, age, degree level, and years in current position. The mean scores from the selected groups were analyzed to measure if any variance (significant differences) existed among group's practices and policies of student affirmative action.

Instrument

The survey instrument used was designed by its original developer to measure the use of student affirmative action practices and policies. The survey questions reflected six areas of affirmative action that included; (1) strict scrutiny analysis, (2) race-targeted financial aid analysis, (3) race-neutral alternatives, (4) special support services, (5) admission status, and (6) affirmative action program tailoring. Based on Cronback's ", the reliability coefficient was found to have a respectably high alpha coefficient of .8126 signifying a strong internal consistency. The survey used a

five-point Likert-type scale ranging from 5 – almost always to 1 – never.

Results

Since a response rate of 50 percent to the survey was desired, a follow-up process for response was utilized. This mailing yielded a return of 39 out of 149 for a response rate of 52 percent. Data from the returned surveys were coded and subsequently entered for analysis.

The majority of survey respondents were male, white, and nearly two thirds of respondents were above 40 years of age. There was little racial diversity among the sample in this study. Seventy-two percent of respondents were white, 23% were African American, and 5% were Hispanic. The respondents were approximately equally distributed based on institutional department; 29% represented central administration, 28% represented student support services, 24% represented admissions, and 19% represented the financial aid department. From this number 64% listed their position as Departmental Director/Assistant Director, and 36% of the respondents held the position of chancellor/president or vice chancellor/vice-president. Specific frequencies for all demographics are shown in Table 1.

Table 1
Demographic Characteristics of the Sample

Variable	Number	%
Sex		
Male	72	65
Female	48	45
Age		
25 - 30	7	5.9
30 - 35	21	17.7
40 - 45	14	11.8
45 - 50	9	7.6
50+	46	38.9
Ethnicity/Race		
Caucasian	85	72.1
African American	27	
Hispanic ~	6	5.1
Department		
CEO	35 ·	28.9
Student Support	34	28.0
Admissions	29	23.9
Financial Aid	23	19.1
Position		
Director/Assistant Director	78	64.5
President/Chancellor	13	10.7
Vice President/Vice Provost		24.8
Number of Years in Position		
Less than Five	43	35.5
Five to Ten	39	32.2
Ten or more	39	32.2

Demographic information regarding the institutional characteristics are presented in Table 2. Sixty-five, or 54% of respondents listed their institution as public, with 46% responding as representing private institutions. Over half, 56%

320

responded as being moderately selective institutions, 25% as selective, and 19% as having open admission status. As for the institutional size, 47% responded as having under 5,000 students, 26% represented institutions between 5,000 to 10,000 students, and 25% of respondents represented a student body of over 10,000. Forty-seven percent stated having an African-American student population under 5%, thirty-eight percent responded as having between 5% to 10%, ten percent answered with having between 10% to 15%, and under eight percent responded having an African-American student body above 15%. Specific frequencies for characteristics of the institutions are shown in Table 2.

Table 2
Characteristics of the Institutions

Variable	Number	Percentage
Institution		
Public	65	53.7
Private	56	46.3
Admission Status		
Open	23	19.5
Moderate Selective	66	55.9
Selective	29	24.5
Size of Institution		
< 5,000	56	48.3
5,000 to 10,000	29	25.0
> 10,000	31	26.7
Percent of		
African American Students		
< 5	53	44.5
5 - 10	45	37.8
10 - 15	12	10.1
> 15	9	7.6

Note: Due to missing data the Ns for some responses do not sum to 121. The percentages are based on the number of responses provided; in some cases this was less than 121.

The dependent variables in this study were the perceived levels of affirmative action policies and practices used by the subjects in six areas of practices and policies. This was obtained from subject's responses to the survey questions. Based on the construction of the survey instruments scale, a high mean (3.0 >) indicated a greater perceived level of use in applying student affirmative action practices and policies. A low mean (< 3.0) represents a perceived lower level of use in student affirmative action practices and policies. The mean for all questions combined, total $M = 3.21$. The sample's responses based on individual questions are represented in Table 3.

Table 3
Descriptive Statistics for Individual Questions

Variable	Mean	SD	Var.
Q1	4.39	.93	.868
Q2	3.47	1.43	2.058
Q3	3.68	1.28	1.628
Q4	2.69	1.41	1.978
Q5	2.94	1.38	1.917
Q6	2.52	1.56	2.437

Q7	2.87	1.52	2.322
Q8	3.15	1.43	2.056
Q9	3.68	1.34	1.804
Q10	2.94	1.85	3.412
Q11	3.37	1.33	1.764
Q12	3.60	1.17	1.378
Q13	3.68	1.40	1.969
Q14	2.64	1.57	2.471
Q15	2.13	1.31	1.714
Q16	2.70	1.44	2.069
Q17	3.03	1.42	2.008
Q18	3.16	1.32	1.755
Q19	3.77	1.42	2.006
Q20	3.89	1.36	1.841
Q21	4.47	1.02	1.038
Q22	2.26	1.22	1.488

The survey questions were grouped into six areas of student affirmative action practices and policies. The six areas included a strict scrutiny analysis, race-targeted financial aid analysis, race-neutral alternatives, special support services, admission program analysis, and affirmative action program tailoring. This information was obtained from subject's responses to survey questions listed in Table 3. The groupings according to question number are as follows: strict scrutiny analysis - Q1, Q2, and Q3; race-targeted financial aid - Q4, Q5, Q6; race-neutral alternatives - Q7, Q8, Q9; special support services - Q10, Q11, Q12; admission program analysis - Q13, Q14, Q15; and affirmative action program tailoring - Q16 thru Q22. The perceived levels of student affirmative action practices and policies are listed respectively in Table 4.

Table 4
Descriptives from the Six Survey Areas

Variable Grouping	Mean	SD	N
Strict Scrutiny Analysis			
Q1	4.39	.93	119
Q2	3.47	1.43	115
Q3	3.68	1.28	116
Total	3.21	.7266	120
Race Targeted Financial			
Aid Analysis			
Q4	2.69	1.41	111
Q5	2.94	1.38	114
Q6	2.52	1.56	120
Total	2.68	1.174	120
Race Neutral			
Alternatives Analysis			
Q7	2.87	1.52	116
Q8	3.15	1.43	117

Q9	3.68	1.34	117
Total	3.26	.9904	11
Special Support Services			
Analysis			
Q10	2.94	1.85	119
Q11	3.37	1.33	120
Q12	3.60	1.17	119
Total	3.31	1.18	120
Admissions Program			
Analysis			
Q13	3.68	1.40	120
Q14	2.64	1.57	118
Q15	2.13	1.31	116
Total	2.86	1.17	120
Affirmative Action Program			
Tailoring Analysis			
Q16	2.70	1.44	113
Q17	3.03	1.42	116
Q18	3.16	1.32	111
Q19	3.77	1.42	115
Q20	3.89	1.36	117
Q21	4.47	1.02	118
Q22	2.26	1.22	114
Total	3.36	.8240	120

Statistics for Scale			
N	Mean	SD	Var.
121	3.21	.528	.7266

Note: Due to missing data the Ns for some responses do not sum to 121. The percentages are based on the number of responses provided; in some cases this was less than 121.

Responses to Survey as Related to Subject Groupings

Table 5 presents the mean and standard deviation of individual's responses and grouped according to gender. The mean difference between the two groups is minimal.

Table 5
Descriptives based on Gender

Gender	Mean	Std. Dev.	Cases
Male	3.21	.7115	72
Female	3.23	.7372	48

Table 6 presents the mean and standard deviation of individual's responses and grouped according to age.

Table 6
Descriptives based on Age

Age	Mean	Std. Dev	Cases
25 - 30	3.51	.9292	7
30 - 35	3.19	.6845	21
35 - 40	3.19	.5439	14
40 - 45	3.37	.5074	21
45 - 50	3.30	.3884	9
> 50	3.02	.8541	46

Note: Due to missing data the Ns for some responses do not sum to 121. The percentages are based on the number of responses provided; in some cases this was less than 121.

The largest difference of means between groups based on age were between the youngest professional age group (25 - 30), and the above 50 age grouping for professionals. Table 7 represents the mean and standard deviation of individual's responses, and subsequently grouped according to ethnicity/race.

Table 7
Descriptives based on Ethnicity/Race

Ethnicity/Race	Cases	Mean	Std. Dev.
Caucasian	85	3.14	.6942
African American	27	3.28	.7660
Hispanic	6	3.28	.7019

Note: Due to missing data the Ns for some responses do not sum to 121. The percentages are based on the number of responses provided; in some cases this was less than 121.

Professionals grouped according to their ethnicity showed only a minimal mean variance. The mean average between African American and Hispanic professionals were identical. Furthermore, the mean average between the previously observed groups when compared to white professionals was minimal. Table 8 presents the mean and standard deviation of individual's responses grouped based on their respective department within the institution.

Table 8
Descriptives based on Department

Department	N	Mean	Std. Dev.
CEO	35	3.17	.7550
Student Support	34	3.08	.8230
Admissions	29	3.21	.5885
Financial Aid	23	3.46	.6703

Note: CEO represents those professionals working in central administration.

Institutional Financial Aid Professionals scored the highest mean of the four group represented. Overall there was only a modest variance between group mean scores based on the professionals institutional department. Table 9 presents the mean and standard deviation of individual's responses and grouped according to their respective position within the institution.

Table 9
Descriptives based on Position

324

Position	N	Mean	SD
President/Provost	13	2.93	.7888
Vice President/Vice Provost	30	3.20	.8138
Director/Assistant Director	78	3.21	.6788

The difference in mean scores of the three groups was relatively small. Two of the groups represented were separated by a score of .01. Presidents and Provost had the lowest group mean (M = 2.93). Overall there was only a small variance between the three groups.

Table 10 presents the mean and standard deviation of individual's responses and grouped based on the number of years in current position.

Table 10
Descriptives based on Number of Years in Current Position

No. of Years	N	Mean	SD
< than 5	43	3.39	.6429
Five to Ten	39	3.28	.6152
Ten >	39	2.94	.8460

The professionals were closely distributed when grouped according to their number of years at current position. Professionals with more than ten years in current position recorded the lowest mean score (M = 2.94). Consequently, professionals with the least number of years in current position recorded the highest mean score (M = 3.39).

Data Analysis

The Statistical Package for the Social Sciences (SPSS) was used to compute the analysis. Following the collection of data the major statistical analysis used was an analysis of variance (One-Way ANOVA). Additionally, the researcher in this study analyzed selected hypothesis using a t-test. Hypothesis 2, 4, 5, 6, 7, and 8 were tested using the One-Way ANOVA. The reason for this was to test multiple groups (variables) for comparison. Hypothesis 1 and 3 were tested using the t-test. Hypothesis 1 and 3 compared two distinct groups. The mean scores for the selected samples were compared to measure the degree of variance between groups. The .05 level of significance was used for all statistical analysis. This section is organized into eight categories based on the hypotheses tested in this study.

Analysis Between Professionals within Public and Private Missouri Institutions

There was a significant difference between the type of institution, public or private at the .05 level, t (119) = 4.26, p < .001. Based on the respondents perceived level of use in student affirmative action practices and policies respondents representing private institutions perceived level of student affirmative action was less (M = 2.92, SD = .678) than respondents representing public institutions (M = 3.45, SD = .681). (See Table 11)

Applying the t-test for independent samples resulted in rejecting the null hypothesis for professionals grouped according to institution (public or private). This finding suggested that the independent variable had an effect on the dependent variable. Institutional personnel do differ significantly in their perceived level of use in student affirmative action practices and policies based on the institution being public or private.

Table 11
t-test for Independent Samples

Equal variances assumed	df	t	p
	119	4.262	< .001

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

To follow up on the differences between the groups, an analysis of variance between the six areas of student affirmative action practices and type of institution was performed. The significant differences between groups fell into three categories, special support services, admission program analysis, and affirmative action program tailoring. (See

Table 12).

Analysis revealed that the significant differences occurred between groups in the following areas; special support services, admission program analysis, and affirmative action program tailoring. This difference was significant at the $p < .05$ level.

Table 12
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
Strict Scrutiny	1	118	3.32	.071
Race Targeted Financial Aid	1	118	2.81	.096
Race Neutral	1	117	.690	.408
Special Support Services	1	118	8.63	.004
Admissions Program	1	118	9.19	.003
Narrow Tailoring	1	118	20.29	<.001

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Analysis Between Professionals Grouped by Ethnicity

The three groups analyzed consisted of Caucasian, African American, and Hispanic. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the level of use in student affirmative action practices and policies based on ethnicity. There was no significant difference between the subjects grouped according to ethnicity at the .05 level, $F(2, 115) = .455, p = .05$. (See Table 13)

Applying the analysis of variance resulted in accepting the null hypothesis for participants grouped according to ethnicity. This finding suggested that institutional personnel grouped according to ethnicity do not differ significantly in their perceived level of use in student affirmative action practices and policies.

Table 13
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
Ethnicity	2	115	.455	.636

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Analysis Between Participants Based on Gender

To determine if a significant difference existed between professionals grouped according to gender, a t-test was conducted. As illustrated in Table 14, there was no significant difference between the groups based on gender at the .05 level, $t(116) = -.054, p = .957$. Based on the respondents perceived level of use in student affirmative action practices and policies, professionals grouped according to gender perceived level of student affirmative action was not significant. For male professionals ($M = 3.2319, SD = .7272$), and for female professionals ($M = 3.239, SD = .7372$).

Applying the t-test for independent samples resulted in accepting the null hypothesis for participants grouped according to gender. This finding suggested that institutional personnel grouped according to gender do not differ significantly in their perceived level in the use of student affirmative action practices and policies.

Table 14
t-test for Independent Samples

Variable	t	df	p
Gender	-.054	116	.957

Analysis Between Professionals Grouped by Institutional Admission Criteria

The three groups analyzed represented institutions having open admission status, being moderately selective, and selective in criteria for admission. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the level of use in student affirmative action practices and policies based on admission status. There was no significant difference between the subjects grouped according to institutional admissions requirements at the .05 level, $F(2, 115) = 2.42, p = .093$. (See Table 15)

Applying the analysis of variance resulted in accepting the null hypothesis for participants grouped according to the admission status of their institution. This finding suggested that institutional personnel do not differ significantly in their perceived level of use in student affirmative action practices and policies based on the institutional admission status.

Table 15
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
Admission Status	2	115	2.42	.093

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Analysis Between Professionals Grouped by Size of Institution

The three groups analyzed represented institutions having a student body enrollment of under 5,000, between 5,000 to 10,000, and above 10,000. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the perceived level of use of student affirmative action practices and policies based on the institutional student body enrollment. There was a significant difference between the professionals grouped according to institutional size at the .05 level, $F(2, 113) = 13.46, p < .001$. (See Table 16)

Table 16
ANOVA Summary Table

Variable		SS	df	MS	p
Institutional Size	Between Groups	11.86	2	5.933	<.001
	Within Groups	49.81	113	.441	
	Total	61.68	115		

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Applying the analysis of variance resulted in rejecting the null hypothesis for participants grouped according to the size of their respective institution. This finding suggest that institutional personnel do differ significantly in their perceived level of use in student affirmative action practices and policies based on the institutional size.

Since the computed F value was significant, Tukey's HSD post hoc test was conducted to determine which groups significantly differed in their perceptions toward the use of student affirmative action policies and practices. Results are listed in Table 17.

Table 17
Dependent Variable: Total; Tukey HSD

Variable	Variable (J) Size	Mean Diff.	Std Error	p

(I) Size		(I) - (J)		
< 5,000	5,000 - 10,000	.4081	.1519	<.022
	10,000+	.7575	.1486	<.001
5,000 to	< 5,000	.4081	.1519	<.022
10,000	10,000+	.3494	.1715	.108
10,000+	< 5,000	.7575	.1486	.001
	5,000 to 10,000	.3494	.1715	<.108

The mean difference is significant $p < .05$ level.

Post hoc analysis using Tukey's HSD test was computed at the .05 level. Analysis revealed that the less than 5,000 institutional group differed significantly ($M = 2.89$, $SD = .6377$) from the other two groups. The 5,000 to 10,000 group ($M = 3.29$, $SD = .7986$), and the 10,000+ group ($M = 3.64$, $SD = .5653$), revealed no significant difference at the .05 level. Institutional size does have an effect on personnel's perception of levels in the use of student affirmative action practices and policies.

Analysis Between Professionals Grouped by the Number of Years in Position

The three groups analyzed represented professionals years of service in current position at their respective institutions. The professionals were grouped accordingly; less than five years of service, five to ten years of service, and, above ten years of service. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the perceived level of use in student affirmative action practices and policies based on professional years in position. There was a significant difference between the professionals grouped according to years in position at the .05 level, $F(2, 118) = 4.42$, $p = .014$. (See Table 18)

Table 18
ANOVA Summary Table

Variable		SS	df	MS	F
Years in Position	Between Groups	4.417	2	2.209	.014
	Within Groups	58.939	118	.499	
	Total	61.68	120	120	

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Applying the analysis of variance resulted in rejecting the null hypothesis for participants grouped according to the number of years in position. This finding suggested that institutional personnel do differ significantly in their perceived level of use in student affirmative action practices and policies based on the institutional size.

Since the computed F value was significant, Tukey's HSD post hoc test was conducted to determine which groups significantly differed in their perceptions toward the use of student affirmative action policies and practices. Results are listed in Table 19.

Table 19
Dependent Variable: Total; Tukey HSD

Variable (I) Size	Variable (J) Size	Mean Diff. (I) - (J)	Std Error	p
< five	five to ten	.114	.1563	.756
	> ten	.4499	.1563	<.013
five to ten	< five	-.1114	.1563	.756
	ten>	.3385	.1600	.091

The mean difference is significant $p < .05$ level.

Post hoc analysis using Tukey's HSD test was computed at the .05 level. Analysis revealed that professional with less than five years differed significantly ($M = 3.39$, $SD = .6429$) from the professionals with more than ten years in their current position ($M = 2.94$, $SD = .8460$). The professionals with five to ten years ($M = 3.28$, $SD = .2.94$), and the professionals with more than ten years in their current position ($M = 2.94$, $SD = .8460$), revealed no significant difference at the .05 level. Furthermore, the professionals with less than five years revealed no significant difference when compared to the professionals with five to ten years of experience in their respective positions. The number of years in position does have an effect on personnel's perception of levels in the use of student affirmative action practices and policies.

Analysis Between Professionals Grouped by Institutional Position

The three groups analyzed represented institutional presidents/chancellors, vice presidents/associate chancellors, and departmental directors. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the level of use in student affirmative action practices and policies based on institutional position. There was no significant difference between the subjects grouped according to institutional position at the .05 level, $F(2, 118) = 1.14$, $p = .323$. (See Table 20)

Table 20
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
Position	2	118	1.141	.323

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Applying the analysis of variance resulted in accepting the null hypothesis for participants grouped according to their position within the institution. This finding suggested that institutional personnel do not differ significantly in their perceived level of use in student affirmative action practices and policies based on the institutional admission status.

The researcher for this study also analyzed professionals perceived levels of the use in student affirmative action based on their respective departments. The four groups analyzed represented the department of admissions, financial aid, student support services, and central administration. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the perceived level of use in student affirmative action practices and policies based on professionals grouped by department. There was no significant difference between the subjects grouped according to institutional position at the .05 level, $F(3, 117) = 1.29$, $p = .278$. (See Table 21)

Table 21
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
Department	3	117	1.298	.278

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the number of responses provided; in some cases this was less than 121.

Applying the analysis of variance resulted in accepting the null hypothesis for participants grouped according to their position within the institution. This finding suggested that institutional personnel do not differ significantly in their perceived level of use in student affirmative action practices and policies grouped according to institutional department.

Analysis Between Professionals Grouped by Institutional Percent of African American Students

The four groups analyzed represented institutions having an African American student body enrollment of less than 5 percent, within 5 percent to 10 percent, between 10 percent and 15 percent, and above 15 percent. After determining that the data met the assumption of homogeneity of variance, a One-way ANOVA was calculated to determine if there was a significant difference in the perceived level of use of student affirmative action practices and policies based on the percent of African American students within the institution.

There was a significant difference between the professionals grouped according to institutional percent of African American students at the .05 level, $F (3, 115) = 13.103, p < .001$. (See Table 22)

Table 22
ANOVA Summary Table

Variable	df_{bg}	df_{wg}	F	p
% of African American Students	13.103	118	10.02	.001

Note: Due to missing data the Ns for some responses do not sum to 121. The df is based on the

number of responses provided; in some cases this was less than 121.

Applying the analysis of variance resulted in rejecting the null hypothesis for participants grouped according to the percent of African American students within the institution. This finding suggested that institutional personnel do differ significantly in their perceived level of use in student affirmative action practices and policies based on the percent of African American students.

Since the computed F value was significant, Tukey's HSD post hoc test was conducted to determine which groups significantly differed in their perceptions toward the use of student affirmative action policies and practices. Results are listed in Table 23.

Table 23
Dependent Variable: Total; Tukey HSD

Variable I % of African American Students	Variable J % of African American Students	Mean Diff. (I) - (J)	Std Error	p
below 5 %	5 - 10 %	-.4341	.1338	<.008
	10 - 15 %	-.5432	.2110	.054
	15 % above	.6968	.2379	<.021
5 - 10 %	below 5 %	.4341	.1338	<.008
	10 - 15 %	-.1090	.2144	.957
	15 % above	1.130	.2410	<.001
10 - 15 %	below 5 %	.5432	.2110	.054
	5 - 10 %	.1090	.2144	.957
	15 % above	1.240	.2910	<.001
15 % above	below 5 %	-.6968	.2379	<.021
	5 - 10 %	-1.130	.2410	<.001
	10 - 15 %	-1.240	.2910	<.001

Post hoc analysis using Tukey's HSD test was computed at the .05 level. Analysis revealed that professional with an African American student population of above 15 percent differed significantly ($M = 2.35, SD = .5778$) from the professionals representing the additional three groups. Furthermore, the professionals with an African American student population less than 5 percent ($M = 3.04, SD = .6876$) showed a significant difference when compared to the professionals with a 5 to 10 percent ($M = 3.48, SD = .6870$) African American student population in their respective institutions. Professionals with an African American student population between 10 - 15 percent ($M = 3.58, SD = .4334$), and professionals representing groups with less than 5 percent ($M = 3.04, SD = .6876$), and between 5 - 10 percent ($M = 3.48, SD = .6870$) indicated no significant difference at the .05 level. In some cases the percent of African American enrollment at an institution does have an effect on personnel's perception of levels in the use of student affirmative action practices and policies.

Of the eight null hypotheses analyzed in this study, the researcher accepted the null for hypotheses two, three, four, and seven. These hypotheses accepted are as follows:

- Hypothesis two: There are no significant differences regarding perception toward affirmative action practices between participants grouped by ethnicity. The analysis of the data revealed no significant differences existed

between participants based on their ethnicity.

- Hypothesis three: There are no significant differences regarding perception toward affirmative action practices between participants grouped by gender. The analysis of the data revealed no significant differences existed between participants based on their gender.

- Hypothesis four: There are no significant differences regarding perception toward affirmative action practices between Missouri institutions based on admission classification. The analysis of the data revealed no significant differences existed between Missouri institutions based on admission classification.

The four hypotheses rejected by the researcher included hypothesis one, five, six, and eight. These hypotheses rejected by the researcher are as follows:

- Hypothesis one: There are no significant differences regarding perception toward affirmative action practices among Missouri personnel based upon the institution being public or private. The analysis of the data revealed that a significant difference existed among Missouri personnel based on their institution being public or private.

- Hypothesis five: There are no significant differences regarding perception toward affirmative action practices between Missouri institutions based on size of institution. The analysis of the data revealed that a significant difference existed among Missouri institutions based on their student population (i.e., size of institution).

- Hypothesis six: There are no significant differences regarding perception toward affirmative action practices between participants based on number of years in position at institution. The analysis of the data revealed that a significant difference existed among Missouri personnel based on their number of years in current position within their respective institution.

- Hypothesis eight: There are no significant differences regarding perception toward affirmative action practices currently used by Missouri personnel when institutions are grouped according to the percentage of African American student enrollment. The analysis of the data revealed that a significant difference existed among Missouri personnel based on their percentage of African American students.

Conclusions

Analysis of the data suggested that Missouri personnel are aware of the judicial scrutiny by the courts in the administering of student affirmative action. However, according to responses personnel in Missouri institutions are not consistent in critiquing their student affirmative action practices and policies. Overall, student affirmative action program objectives serve two purposes: (a) remedy the present effects of past discrimination, and (b) to advance campus diversity.

Concerning financial aid, Missouri institutions occasionally used race/ethnicity awards to attract students of color to their respective institutions. Provided race/ethnicity awards are used, the application of statistical data to support race/ethnicity awards are used occasionally by Missouri institutions. This finding contradicts with the fact that Missouri personnel are mindful of the judicial scrutiny by the courts in the administering of student affirmative action. Race neutral alternatives, such as socioeconomic statuses are currently being administered in place of race/ethnicity financial awards at Missouri institutions.

The issue of student diversity currently is a concern for Missouri institutions. Designed programs for retention, separate departments such as Minority Affairs Offices, and the identification of faculty mentors for African American students are supported by Missouri institutions. Overall, Missouri institutions actively target and recruit prospective African American students for the specific purpose of campus diversity. The data revealed little indication that Missouri institutions are currently administering special allotments for admission. Missouri institutions did not suggest that separate pools, subcommittees, and separate cutoff scores were a part of current practice and policy.

Overall, Missouri institutions have taken steps to reduce the impact of currently used affirmative action practices on students not eligible for participation. An overwhelming majority of Missouri institutions use a single process for assessing all applicants for admission, without the reliance of a quota system. The recent Hopwood decision revealed limited impact on the decisions regarding professionals use of student affirmative action at Missouri institutions.

There are several authors and researchers within the context of higher education addressing questions regarding perceptions toward student affirmative action (Bowen & Bok, 1998). The United States Department of Education has provided guidelines for those in higher education to assist in developing permissible student affirmative action policies. However, it appears that most, if not all, of these policies are not from the perspective of professionals in the field.

331   BEST COPY AVAILABLE

The results of this study suggest a positive picture of student affirmative action practices and policies used by Missouri personnel. The overall mean score for support in diversifying Missouri institutions was relatively high. Perceived differences among groups were at a minimum. In analyzing the perceived difference between public and private Missouri institutions revealed a higher overall mean score for public institutions. This was expected due to the fact public institutions must comply with federal guidelines for affirmative action as set by the Office of Federal Contract Compliance Programs (OFCCP), and the statement released by U. S. Secretary of Education Richard W. Riley in response to the passage of Proposition 209 (United States Department of Education press release, March 1997). Furthermore, a higher mean level for public institutions may reflect diversity initiatives taken by the Missouri Coordinating Board for Higher Education in the late 1980s, and early 1990s.

Although the majority of all survey respondents were male (65%), and Caucasian (72%), this group appeared to have no perceived difference in their level of use in student affirmative action. Overall, their responses were similar to those perceived levels by African American and Hispanic professionals. Clearly, their perceptions of student affirmative practices and policies were positive. Similarly the groups compared closely with gender used as a variable in this study. Women (35%) respondents displayed no difference in their analyzed perceived levels of student affirmative action when compared with male professionals.

There are three levels of criteria for universities in selecting their student body based on admission requirements. According to Cross and Slater (1997) the authors' assessments suggested that if standardized tests become the single norm in admission decisions, African-American enrollment at some institutions will drop dramatically. Most of the respondents represented moderately selective institutions (56%), with professionals representing open admission 20 percent, and selective as 24 percent. Overall, their responses were similar toward perceived levels of student affirmative action. Interesting the data revealed selective institutions as having a slightly higher level in the use of student affirmative action. Although the researcher did not acquire individual institutional admission requirements, this finding suggested that admission criterion does not affect professionals perceptions toward policies and practices in student affirmative action.

Levels of perceptions in student affirmative action practices and policies were higher in institutions with an enrollment of more than 10, 000 students. Concerning student support services, and strict scrutiny analysis, institutions with more than 10, 000 students had noticeable higher levels of perceived use in student affirmative action. The researcher can only offer two assumptions for this attained higher level. In the area of strict scrutiny, the majority of lawsuits against student affirmative action practices have been directed toward large state institutions (*Regents of Univ. of Cal. V. Bakke*, 1978; *Podberesky v. Kirwan*, 1994; *Texas v. Hopwood*, 1996). Secondly, there may be greater state allocations (i.e., funding) available for these institutions toward recruitment, and retention of African American and other minority groups.

Although most of the respondents had less than five years of experience in their current position (35%), the groups were closely distributed. This group also displayed a higher level of perceived use in student affirmative action practices and policies. Clearly, the less number of years in position appeared to have an impact on the perceptions of this group. This was an interesting and puzzling finding since the two areas of significant difference represented the financial aid analysis, and race neutral alternatives. The researcher expected this variable to have little difference between the groups. This relationship may have been attributed more toward a greater responsibility of professionals following their institutional practices based on position assurance. Professionals with more seniority may feel a greater sense of security within the institution due to longevity or tenure. This was one variable the researcher did not account for in this study. However, seniority and tenure could have an impact on perceptions toward institutional practices and policies. This statement would account for the differences in these two areas of practices and policies for professionals with less than five years in their current position.

For professionals position within their institution, the data revealed no significant difference between groups. Directors and Assistant Directors displayed a slightly higher group level of perceptions toward student affirmative action practices and policies. This higher level corresponds with this group of professionals since they are more actively involved in the conduction of student affirmative action policies and practices.

Accordingly, when professionals were analyzed based on their department within the institution, the data revealed no difference between groups. Understandably, since other variables analyzed were similar for perceived levels, analysis presented great consistency among the four departments represented by Central Administration, Admissions, Financial Aid, and Student Support Services. Overall, the groups exhibited a perceived level favorable toward student affirmative action.

The final variable analyzed in this study investigated perceived levels toward student affirmative action based on the percentage of African American students. Post hoc analysis revealed that professionals with an African American student population of above 15 percent differed significantly from the professionals in the other three groups. This difference may be attributed to the fact that institutional personnel with less than 15 percent are more aware of their need to increase campus diversity. Therefore, these groups' levels of perceptions were greater than those exhibiting a higher percentage of African American students on campus. This would explain the higher mean level for groups with less than 15 percent African American student representation. The second explanation is that those institutions with less than 15 percent represent areas with minimal community diversity. Therefore, the need for student affirmative action policies and practices becomes more urgent. In some cases, the percent of African American enrollment at an

institution does have an effect on personnel's perception of levels of use of student affirmative action practices and policies.

References

Alexander, Lamar. United States Secretary of Education, Press Release Statement, March 20, 1991. Washington D.C.

Bolden, V. A., Goldberg, D. T., & Parker, D. D. (1999). Affirmative action in court: The case for optimism. *Journal of Equity & Excellence in Education, 32* (2), 24-30.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long term consequences of considering race in college and university admissions.* Princeton, NJ: Princeton University Press.

Bunzel, J. H. (1995). The California Civil Rights Initiative: The debate over race, equality and affirmative action. *Vital Speeches of the Day, 61* (17), 530-533.

Burd, S. (1998). House votes down proposal to bar racial preferences in admissions. *Chronicle of Higher Education, 44* (36), A35.

Cross, T., & Slater, R. (1997). Why the end of affirmative action would exclude

all but a very few Blacks from America's leading universities and graduate schools. *Journal of Blacks in Higher Education, 17*, 8-17.

Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application.* (6th ed.). N J: Merrill Prentice Hall.

Henry, A. R. (1998). Perpetuating Plessey v. Ferguson and the dilemma of Black

access to public higher education. *Journal of Law and Education, 27* (1), 47-71.

Hopwood v. Texas, 78 F.3d 932 (5th Cir.), cert denied, 116 S.Ct. 2581 (1996).

Hopwood v. Texas, 861 F. Supp. 551 (W.D. Tex. 1994), rev'd, 78 F.3d 932 (5th Cir.), cert. Denied, 116 S.Ct. 2581 (1996).

Jones, T. (1998). Life after Proposition 209. *Academe, 84* (4), 22-28.

Kurlaender, M., & Orfield, G. (1999). In defense of diversity: New research and evidence from the University of Michigan. *Journal of Equity & Excellence in Education, 32* (2), 31-35.

Missouri Coordinating Board for Higher Education. (1988). *Challenges and opportunities: Minorities in Missouri higher education.* Coordinating Board for Higher Education (CBHE), Jefferson City, MO.

Office of Federal Contract Compliance Programs, Title 41 C.F.R. Section 60-2.1 (1978).

Office of Federal Contract Compliance Programs, Title 41 C.F.R. Section 60-2.13 (1978).

Podberesky v. Kirwan, 38 F.3d 147 (4th Cir. 1994), cert denied, 115 S.Ct. 2001 (1995).

Regents of Univ. of Cal. V. Bakke, 438 U.S. 265 (1978).

Regents of Univ. of Cal. V. Bakke, 438 U.S. 265, 311-315 (1978) (opinion of Powell, J.)

About the Author

Alfred R. Cade, Jr.
School of Education
Missouri Southern State College
Joplin, MO 64801-1595

Email: cade-a@mail.mssc.edu

Al Cade, Ed.D., is currently the Assistant to the Dean for the School of Education at Missouri Southern State College. He is a former president for the Missouri Association for Blacks in Higher Education (MABHE). His research and scholarship interests include policies in K-12 and higher education. As an Assistant Professor in the Department of Teacher Education, his teaching specialties include multicultural education, diversity, and social policies.

334

adrianacosta@compuserve.com

Teresa Bracho (México)
Centro de Investigación y Docencia Económica-CIDE
bracho dis1.cide.mx

Ursula Casanova (U.S.A.)
Arizona State University
casanova@asu.edu

Erwin Epstein (U.S.A.)
Loyola University of Chicago
Eepstein@luc.edu

Rollin Kent (México)
Departamento de Investigación Educativa-
DIE/CINVESTAV
rkent@gemtel.com.mx   kentr@data.net.mx

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

felix.angulo@uca.es

Alejandro Canales (México)
Universidad Nacional Autónoma de México
canalesa@servidor.unam.mx

José Contreras Domingo
Universitat de Barcelona
Jose.Contreras@doe.d5.ub.es

Josué González (U.S.A.)
Arizona State University
josue@asu.edu

María Beatriz Luce (Brazil)
Universidad Federal de Rio Grande do Sul-UFRGS
lucemb@orion.ufrgs.br

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

Education Policy Analysis Archives

School-Based Management: Views from Public and Private Elementary School Principals

Mary T. Apodaca-Tucker
New Mexico State University

John R. Slate
University of Texas at El Paso

Abstract

In this study, we analyzed the principal questionnaire contained in the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) database regarding the extent to which school-based management was reported as having been implemented differently by public and by private elementary school principals. Statistical analyses indicated many differences in the degree of influence reported to be present on the part of principals, parents, and other groups on important decisions made at schools. Differences in school-based management between our public and private elementary school principals were linked to the extant literature. Moreover, recommendations for further research were discussed.

In 1991, the Texas Education Agency directed schools to form school-based decision-making committees. Other states in this nation have created similar mandates to reform their schools. The ultimate purpose of all decision-making in schools is to achieve the state's educational goals of equity and excellence for all students. Committees also served as advisory councils to the principal. Shared decision-making (SDM) committee was to include parents, teachers, administrators, and community representatives. Because of the increased local autonomy and accountability that is created through SDM, increased student achievement has been cited as a positive outcome of SDM (TEA, 1992). Strong leadership by school principals has also been supported by the Department of Education in the report entitled Turn Around Low-Performing Schools (U.S. Department of Education, May 1998). Limited research, unfortunately, is available about the extent to which school-based management has been implemented across the United States.

Theoretical Basis of the Study

School-based management functions under decentralization, the development of internal resources, and the wide

participation of school members in the decision-making process, which closely accompanies the tenets of critical theory. Livingston, Slate, and Gibb (1999) reported that administrators agree that all stakeholders must be involved in decision-making if the school is to be successful and that teachers possess expertise that is necessary to make important decisions about the school. In addition, Cheng (1996) suggested that SBM assumes a multiplicity of educational goals, a complex and changing educational environment, need for educational reforms, school effectiveness, and the pursuit of quality.

The theory that guides this study is based on the work of two educational researchers: Glickman (1993) and Sergiovanni (1992, 1994, and 2001) as well as researchers Conley (1993) and Schlechty (1997). The framework that guided Glickman's research (1993) consisted of a covenant of teaching and learning that is brought to life using shared governance and action research. A covenant of teaching and learning is a set of belief statements that capture what people associated with a school want students to know and be able to do, the type of instructional practices they believe will bring about these desired results, and a description of how students will demonstrate mastery of the desired skills and understandings. Shared governance is a democratic process that gives all of a school's stakeholders the opportunity to actively participate in bringing their covenant to life. Action research is an information-producing process that provides feedback and guidance as a school works to carry out the terms of its covenant (Glickman, 1993).

Sergiovanni (1992) reported that most educators would agree that leadership is an important component in improving our schools, yet few people are satisfied with leadership practices now in place. Sergiovanni illustrated how creating a new leadership practice, one with moral dimension centered around purpose, values, and beliefs, can transform school from an organization to a community (1994) and inspire the kinds of commitment, devotion, and service that can make our schools great (2001). Sergiovanni agreed with the research by Glickman (1993) by arguing that this new leadership style is importance to legitimizing emotion and getting in touch with basic values and connections with others. Sergiovanni and Glickman both reported in their separate research how collegiality, based on shared work and common goals, leads to a natural interdependence among teachers. When teachers and administrators are motivated by emotional and social bonds, guided by a professional ideal, and feel they are truly part of a community, the guiding principle is no longer what is rewarded occurs, but what is good happens (Conley, 1993; Schlechty, 1997).

Participatory Management

Participative decision-making is not a new concept. Senge (1990) catapulted learning organizations in business into popularity in the 1990s, and he also reported about participative openness. This theory by Senge (1990) about participative management soon became part of the educational reform movement. Researchers, through their literature, illustrated a development in school reform that became known as school-based management.

The concept of school-based management (SBM) and shared decision-making (SDM) basically fell under the theoretical umbrella of participative management. In recent years, it has become a generally accepted belief that people who participate in the decisions that directly affect them are more likely to have a sense of ownership and commitment to the decisions and situations that involve them (Glickman, 1993; Conley, 1993). School systems are beginning to acknowledge the need to reform traditional hierarchical structures and to experiment with participative management styles to meet the needs of students who are falling behind acceptable academic standards (Conley, 1993).

Supposedly the low morale of school employees and the decrease in organizational effectiveness has led many experts in the field of education to recognize the need for organizational and structural change. Educational systems in America have been publicly criticized for being disorganized and having little empathy for the plight of their employees (Conley, 1993). Consequently, it appeared a natural outgrowth that reform related to participative management styles would be a viable consideration to traditional school structures. Teachers who have low morale and a sense of helplessness within their school system would seemingly be less inclined to apply maximum effort or maximum use of their professional capacities when instructing the nation's students (Conley, 1993).

It becomes apparent that participative management is complex in its theoretical structure. Different perceptions of participation may be related to the success or failure of the emergent styles of participative management (SDM, SBM, site-based management) that are currently being considered for implementation or already have been implemented in schools nationwide. How does participative management merge into education?

Shared Decision-Making

Shared decision-making, according to Allan S. Vann's magazine article in *Educational Horizons* entitled "Shared Decision-Making: A Paper Tiger?" (Fall, 1999), is a state mandate that each school have a site-based management committee composed of parents, teachers, and administrators. The purpose of each committee is to engage in shared decision-making to improve student achievement. Consequently, it is left to each local school board, however, to determine each school's precise committee composition, the membership selection process, and the issues that such committees can, and cannot, consider. Researchers have revealed contradicting information from studies on school reform; some researchers reported the advantages, limitations, and components of SDM. Therefore, the following review of the literature on shared decision-making reflects the diversity of information discovered by these writers.

According to Rodriquez (2000), site-based management is implemented in a variety of ways in districts and schools across the United States. One of the reasons for the differences in implementation is a variation in focus. Clune and White (1988) reported that many districts judge SDM as more of a mind set or disposition than a structured system. Malen, Ogawa, and Kranz (1990) stated that the emphasis is more on the spirit of the approach than the details of the arrangement. In addition, they indicated that key parameters are set in place by districts regarding site-based management, but explicit detail of the governance process is left up to the individual school (Hill & Bonan, 1991). In a study conducted by Smith (1993), the conclusion was that districts supplied insufficient clarification of the roles teachers were to play in the decision-making process, and that districts gave little assistance as to how site-based management should be implemented. Ambiguity left by the districts caused teachers to build their own varying definitions of SBM (Smith, 1993). During the investigation of Chicago's school reform conducted by Hess (1991), he found that the first years of site-based management were a time of "informal negotiations" (p. 8) during which shared decision making began to take on meaning.

Rodriguez (2000) reported that investigators have delineated three broad spheres of influence, or domains of site-based management, budgeting, curriculum, and personnel. In addition, goals and organizational structure have been added to these domains by Hill and Bonan (1991). Freedom to develop goals is perhaps one of the most important aspects of self-governing schools. Clark and Meloy (1989) remarked that well-developed goals include the values on which collaborative action can be taken. They also represent agreement on principles according to Hill and Bonan (1991) that aided in the solution of daily matters. Ultimately, control over its mission enables a school to create a distinctive culture and climate that allow it to meet the needs of the local community (Dade County Public Schools, 1989).

Another aspect of site-based management is control over the budget. Autonomy in the sphere of finance is affected in numerous respects, reported Rodriguez (2000). Brown (1990) reported that SDM brings about a change in the manner in which resources are allocated to schools. Therefore, advocates of site-based management called for districts to allocate a lump sum of money to the schools, not to determine how that money is to be spent (Clune & White, 1988). Such an allowance by site-based management permits stakeholders at the school-level to decide how the money will be dispersed. Hannaway (1992) noted that the larger the sum of money allocated to a school, the greater the amount of decentralization.

A key issue that Rodriguez (2000) noted was that the spending of schools' money is the extent to which those schools are able to spend the money as they wish, such as purchasing from venders outside the district. Consequently, schools operating under site-based management generally have greater flexibility regarding how they spend their money and whom they purchase from than schools operating under the traditional model of school governance (Wohlstetter & Buffet, 1991). Hill and Bona (1991) reported that the greater the decentralization in a district, the greater the ability for empowered site-based managed schools to purchase what they need to meet their students' needs.

Closely connected to control over the budget was control over the hiring of school personnel (Rodriguez, 2000). In districts with the least amount of decentralization, hiring was generally left up to the district, whereas districts that were highly decentralized gave nearly full control to their schools over the hiring of staff and faculty (Lindelow, 1981). In successfull site-based managed schools, Lindelow (1981) reported that administrators and teachers, along with community members, select candidates to interview and make a decision, which is sent back to the district for final approval by the school board. Some decentralized districts permitted their schools to choose how they use personnel funding, such as purchasing books or materials or hiring paraprofessionals instead of teachers with the money (Fernandez, 1989). In the most extreme cases of site-based management, control over the hiring of the principal is a decision left up to the site-based decision-making committee (Chapman, 1990).

Another aspect of school-site autonomy was the ability to choose curricula that meet objectives set by the board and district administration (Rodriguez, 2000). School-based curriculum allowed the site-based decision-making committee to determine which instructional materials should be used for instruction (Steffy, 1993). Clune and White (1988) reported that SDM schools make decisions regarding the selection of textbooks, the selection of learning activities and supplemental instructional materials to be used, and determine the nature of alternative programs to be offered in the school.

The more in-depth implementation of site-based management in a district, the more opportunities local communities have to be involved in the selection of theoretical approaches used in the schools (Rodriguez, 2000; Watkins & Lusi, 1989) and in choosing professional development activities that helps teachers meet the needs of the students. In addition, Guthrie (1986) reported that SBM implemented extensively allows for effective monitoring and evaluation of local learning and teaching by the particular school.

A final sphere of influence that Rodriguez (2000) reported was the influence related to site-based management in school organizations. She indicated that decision-making committees are free to change the fundamental delivery of instruction and the traditional set-up of the classroom. Schools expansively implementing site-based management at the elementary level are drastically altering the manner in which students are grouped to form classes, such as changing age and ability combinations (Murphy, 1991). He also argued that secondary schools with widely implemented site-based management have offered alternative instructional programs, core curricula, and outcome-based education to their students.

Numerous authors (Carlson, 1996; Reynolds, 1997) in the literature emphasized the importance of shared decision making training. Whereas teachers are knowledgeable in their own domain, their preparation seldom included a heavy emphasis on collaborative decision-making. Shared decision-making schools used a variety of methods to provide the necessary training, including outside consultants, trains the trainer programs, and the use of specific training methods.

In support of improving schools from within using shared governance, Barth (1990) argued that the personal visions of most school practitioners need no apology. "For certain, they differ in important ways from the lists of desirable school qualities constructed by those outside the schools. But these visions of insiders deserved to be taken as seriously as those of outsiders", (Barth, 1990, p. 177). He illustrated this argument by stating that not one but two tributaries flow into the knowledge base for improving schools: the social science research literature from the academic community and the craft knowledge and vision from the school community. The former is often a mile wide but only an inch deep; the latter is often only an inch wide but a mile deep. Together, they offer remarkable depth and breadth and a fertile meeting place for considering school improvement. Working in a school day after day, or rearing children of their own, entitles school people and parents to have a vision and to introduce that vision into conversations about school reform (Barth, 1990).

As principals struggled with their daily dilemmas of leadership, they sometimes allow themselves daydreams in which their authority is unlimited and they can act without having to plead, lobby, or negotiate with anyone. Yet, for the past decade, many school leaders have willingly participated in a movement that asks them to share their power with teachers and parents. In shared decision-making (SDM), principals collaborated with teachers and sometimes parents to take actions aimed at improving instruction and school climate. In some cases, teachers or parents are formally given a slice of power; more commonly, principals retain their authority but commit themselves to govern through consensus.

After reviewing the literature, it appeared that shared decision-making is still too new to determine its overall effectiveness in schools. Longitudinal studies on the academic achievement of students, school operations, quality of instruction, the perceptions of students, teachers, and administrators must continue to be conducted to determine the effectiveness of SDM as a means for school reform (Herman & Herman, 1994).

Public and Private Schools

Differences in the organization of public and private schools are a focus of school reform discussions. Yet, how different or similar public and private schools really are is not well understood. School sector is not a simple organizational fault line running through the nation's schools. Debates about improving schools often overlook the diversity among private schools, as well as the potential for a high degree of similarity between many public and private schools (Baker, Han, & Keil, 1996; Synder, 1997). Using data from a national sample of secondary schools in the 1990-91 Schools and Staffing Survey, conducted by the National Center for Education Statistics (NCES), examined organizational differences across public and private schools and among private school types (Baker et al., 1996). Overall, the results from researchers indicated considerable organizational variation among different types of private schools and some significant similarities between public schools and some types of private schools. In addition, although private schools tend to have more on-site control of key administrative decisions about teacher hiring, curriculum, and student discipline policies, not all public schools lack this feature. Accordingly, some difference exists in degree of administrative control among types of private schools as well (Baker et al., 1996). Principals reported that on three types of policies, decision-making in private secondary schools is dominated by principals. Private school principals are more likely to have a greater influence over establishing the curriculum than public school principals. However, both private and public school principals have a great deal of influence on hiring (93 versus 84) and disciplinary policy (91 versus 88).

Teachers in only a few schools in both sectors have a great deal of influence on hiring policies. About two-thirds of private schools have important input from teachers into curriculum decisions, compared to just over half of public schools (Baker et al., 1996). School boards had a similar impact on teacher hiring across public and private sectors, but there is variation among private and public school type. Public school boards are more likely to have an influence on curricular and disciplinary policies than private school boards. Therefore, decisions about organizational policy related to the educational functioning of the school tend to be more influenced by on-site personnel in private schools than in public schools. Clear differences are present between the public and private sectors in the governance environment of schools as reported by Baker and his colleagues (1996).

Bryk, Lee, and Holland (1993) indicated that many reform proposals for public schools have looked to the private sector for models to emulate. School choice, small schools, and decentralization decision-making, for example, are among features commonly associated with private education that many have suggested might benefit public schools. The variation that exists is as follows:

- The defining distinction between public and private schools is their different sources of support.
- Private schools provide an alternative for parents who are dissatisfied with public schools or have other reasons for wanting their children to attend a private school.
- Racial and ethnic diversity can enrich the school experiences of students and teachers in many ways; however, a heterogeneous school population creates additional challenges to teachers and administrators, who must be sensitive to different cultural backgrounds.

339

- Differences between public and private school teachers are an important dimension in comparing public and private schools. Public school teachers appear to be more qualified than private school teachers in terms of their education and years of experience. On average, public school teachers receive higher salaries and more benefits than private school teachers. Although teacher attrition tends to be higher in private than public schools, private school teachers were more likely than public school teachers to be highly satisfied with their working conditions (36 % versus 11 %).
- Smaller schools are generally thought to be easier to manage, and to promote a greater sense of community among students and teachers; however, large schools are often more equipped to offer a wider range of academic programs and support services. Private schools, on average, have smaller schools and class sizes than public schools.
- A key aspect of school management is where important decisions are made concerning curriculum, school policies, and classroom practices. Whereas public schools must necessarily take some direction from state departments of education, local school boards, and districts staff, private school teachers and principals are more likely than their public school counterparts to believe they have a great deal of influence, particularly in setting discipline policy and establishing curriculum.
- In the area of teacher evaluation, almost all principals, public and private, thought they had a great deal of influence; however, in a number of other policy areas as discipline, curriculum, inservice training, budgeting, and hiring, private school principals were more likely than public school principals to think that they had a great deal of influence.
- Although crime occurs in and around both public and private schools, public schools have a much greater exposure. In 1993-1994, teachers in public schools were far more likely than private school teachers to report that students' poor attitudes toward learning and negative interactions with teachers were serious problems in their schools. They were also more likely to believe that a lack of parent involvement was a serious problem. Parent accountability and participation in elementary schools may be more associated with the social class of parents than with the private or public character of the school.
- The key aspects of the instructional program at the elementary level are the amount of time spent on core subjects, the teaching methods used in the classroom, and how homework is handled. Public and private schools exhibit both similarities and differences in these areas.
- Public schools provide a wide array of academic support and health-related services, some of which are required by federal and state laws that do not apply to private schools. Most support services are found more often in public schools than private schools (Choy, 1998).

The National Center for Education Statistics conducted a study to determine exactly how public and private schools differ. The data reported many systematic differences, and provided a context in which to consider the debates about the merits of various aspects of public and private schooling. Synder and colleagues (1997) reported that a key aspect of school management is where important decisions are made concerning curriculum, school policies, and classroom practices. Whereas public schools necessarily must take some direction from State Departments of education, local school boards, and district staff, more site-based management and local decision-making are frequently advocated as a means of improving school effectiveness.

- Private school principals (or heads) reported having more influence over curriculum than their private school counterparts.
- In a number of school policy areas, private school teachers and principals are more likely than their public school counterparts to believe that they have a great deal of influence.
- Private school teachers reported having more autonomy in the classroom (Synder, 1997).

In the areas of setting discipline policy and establishing curriculum, in particular, private school teachers in the 1993-94 school year were considerably more likely than public school teachers to think that they had a great deal of influence. Only a relatively small percentage of teachers in either sector were likely that they had a great deal of influence over certain other important policy areas, such as making budget decisions, hiring, and evaluating teachers (Synder, 1997). In contrast, public and private school principals reported they had a great deal of influence in the area of teacher evaluation. However, in a number of other policy areas, discipline, curriculum, in-service training, budgeting, and hiring, private school principals were more likely than public school principals to think that they had a great deal of influence reported Synder and his colleagues (1997). Public school principals share authority for many policy decisions with school boards, district personnel, and State Departments of Education.

The following research question will be addressed in this study. Is there a significant difference in the extent to which school-based management has been reported as having been implemented in public and private elementary schools in the United States?

Methods and Procedures

Sample

In this study, 866 elementary school principals completed the survey. Of these 866 principals, 630 surveys were completed by public elementary school principals and 236 surveys were completed by private elementary school principals. Although not analyzed by category within private elementary schools, 105 private schools were Catholic, 75 were Other Religious category, and 56 were Other private. Information on the survey was also present regarding

school characteristics such as region and location and student enrollment. Regarding school region, 154 were from the Northeast, 228 were from the Midwest, 286 were from the South, and 198 were from the West. In terms of school location, 385 were designated as Central City, 286 were Urban/Large Town, and 195 were Small Town/Rural. Student enrollment ranged from 0-149 students ($\underline{n}$ = 117), 150 to 299 ($\underline{n}$ = 179), 300 to 499 ($\underline{n}$ = 223), 500 to 749 ($\underline{n}$ = 226, and 750 and above ($\underline{n}$ = 121).

Information was also present regarding principal characteristics such as gender and Hispanic ethnicity. Regarding gender, 331 elementary school principals indicated they were male and 517 reported they were female. Of the sample, 36 reported they were of Hispanic ethnicity and 805 indicated they were not non-Hispanic. Other data regarding principal ethnicity was suppressed on the database used herein.

Instrumentation

School administrators, principals, and headmasters were asked to complete self-administered questionnaires during the spring of 1999. They were asked to provide information on the physical, organizational, and fiscal characteristics of their schools and on the school's learning environment and programs. Special attention was paid to the instructional philosophy of the school and its expectations for students.

The questionnaire was an important part of the ECLS-K project and the questionnaire was directed to the school principal. As a result, the questionnaire was divided into nine sections. These sections could have been answered either by the principal or by a designee who was able to provide the requested information. The final two sections requested judgmental evaluations about the school climate and factual information about the principal's background and experience. These last two sections were to be completed by the principal. Some factual questions requested information that was not readily available from school records (the average number of years a limited-English-proficient first grader receives English-as-a-Second-Language services). Informed estimates were acceptable for such questions.

Section 8 focused on school governance and climate. Principals were asked to respond to questions about frequency of classroom observations of kindergarten teachers, staff development, goals and objectives for kindergarten teachers, how decisions are made at their school, the school climate, and what influences the principal's job performance evaluation. Section 9 focused on 10 principal characteristics. The time required to complete this information collection was estimated to average 45 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collected (U.S. Department of Education, April 2000).

Results

The degree to which school-based management had been implemented in public elementary schools in the United States was examined through an analysis of question 67, "We are interested in how decisions are made at your school." Respondents were provided with six decisions: (1) establishing criteria for hiring and firing teachers; (2) selecting textbooks and other instructional materials; (3) setting curricular guidelines and standards; (4) establishing policies and practices for grading and student evaluation; (5) deciding how school discretionary funds will be spent; and (6) planning professional development. Percentages regarding the influence each category of decision maker (i.e., principal or director; teacher organization or individual teachers; parent organization; school board or council; school district office; and school-based management committee) had on each of the decision categories made at their school are reported in Tables 1-6 based on the responses from public and private elementary school principals in the United States.

Table 1
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding The Influence of Decision Makers On The Hiring and Firing of Teachers

Decision Makers	Public	Private
Administrator Input To Hiring/Firing Teachers		
No Influence	6.2	1.5
Some Influence	15.3	6.7
Major Influence	78.5	91.8
Teacher Input To Hiring/Firing Teachers		
No Influence	29.4	49.4
Some Influencex 47.6	39.0	
Major Influence	23.0	11.6
Parent Input To Hiring/Firing Teachers		

No Influence	79.8	82.7
Some Influence	19.2	14.3
Major Influence	1.1	3.0

School Board Member Input To Hiring/Firing Teachers

No Influence	17.0	43.9
Some Influence	22.1	22.9
Major Influence	60.9	33.1

School District Input To Hiring/Firing Teachers

No Influence	8.6	52.5
Some Influence	23.6	22.1
Major Influence	67.8	25.4

School-Based Management Committee Input To Hiring/Firing Teachers

No Influence	58.3	81.7
Some Influence	25.9	10.8
Major Influence	15.9	7.5

Table 2
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding The Influence of Decision Makers On Selecting Textbooks

Decision Makers	Public	Private

Administrator Input On Selecting Textbooks

No Influence	5.6	2.1
Some Influence	48.3	14.9
Major Influence	46.2	83.1

Teacher Input On Selecting Textbooks

No Influence	6.0	3.2
Some Influence	19.4	13.4
Major Influence	74.6	83.3

Parent Input On Selecting Textbooks

No Influence	55.3	67.5
Some Influence	38.5	28.2
Major Influence	6.3	4.3

School Board Member Input On Selecting Textbooks

No Influence	24.6	57.3
Some Influence	36.8	33.1
Major Influence	38.6	9.6

School District Input On Selecting Textbooks

No Influence	11.2	52.0
Some Influence	31.0	29.6
Major Influence	57.8	18.4

School-Based Management Committee Input On Selecting Textbooks

No Influence	38.2	79.3
Some Influence	28.3	16.3
Major Influence	33.5	4.3

Table 3
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding The Influence of Decision Makers On Setting Curricular Guidelines and Standards

Decision Makers	Public	Private
Administrator Input On Setting Curricular Guidelines and Standard		
No Influence	6.5	1.0
Some Influence	39.4	10.4
Major Influence	54.1	88.6
Teacher Input On Setting Curricular Guidelines and Standards		
No Influence	9.4	6.6
Some Influence	37.7	24.2
Major Influence	52.8	69.2
Parent Input On Setting Curricular Guidelines and Standards		
No Influence	46.6	66.3
Some Influence	46.2	31.3
Major Influence	7.3	2.4
School Board Member Input On Setting Curricular Guidelines and Standards		
No Influence	9.6	38.2
Some Influence	27.7	40.8
Major Influence	62.7	21.0
School District Input On Setting Curricular Guidelines and Standards		
No Influence	4.0	41.4
Some Influence	16.0	16.5
Major Influence	80.0	42.1
School-Based Management Committee Input On Setting Curricular Guidelines and Standards		
No Influence	35.4	79.8
Some Influence	34.5	8.5
Major Influence	30.1	11.7

Table 4
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding Establishing Policies and Practices for Student Grading/Evaluation

Decision Makers	Public	Private
Administrator Input On Establishing Policies and Practices for Student Grading/Evaluation		
No Influence	4.8	.5
Some Influence	36.8	11.1
Major Influcnce	58.4	88.4
Teacher Input On Establishing Policies and Practices for Student Grading/Evaluation		
No Influcnce	5.1	4.4
Some Influence	29.2	21.3
Major Influence	65.7	74.3
Parent Input On Establishing Policies and Practices for Student Grading/Evaluation		
No Influence	53.7	77.3
Some Influence	39.0	19.6
Major Influence	7.3	3.1
School Board Member Input On		

Establishing Policies and Practices
for Student Grading/Evaluation

No Influence	12.7	52.3
Some Influence	29.6	29.7
Major Influence	57.7	18.1

School District Input On Establishing
Policies and Practices for Student
Grading/Evaluation

No Influence	5.2	44.7
Some Influence	23.3	17.4
Major Influence	71.5	37.9

School-Based Management Committee
Input Establishing Policies and Practices
for Student Grading/Evaluation

No Influence	38.2	83.0
Some Influence	31.7	8.5
Major Influence	30.1	8.5

Table 5
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding The Influence of Decision Makers On Deciding How School Discretionary Funds Will Be Spent

Decision Makers	Public	Private
Administrator Input On Deciding How		
School Discretionary Funds Will Be Spent		
No Influence	.4	.5
Some Influence	14.0	11.7
Major Influence	85.6	87.8
Teacher Input On Deciding How School		
Discretionary Funds Will Be Spent		
No Influence	11.5	18.4
Some Influence	42.5	54.7
Major Influence	46.0	26.8
Parent Input On Deciding How School		
Discretionary Funds Will Be Spent		
No Influence	40.6	36.3
Some Influence	43.2	43.5
Major Influence	16.2	20.2
School Board Member Input On Deciding How		
School Discretionary Funds Will Be Spent		
No Influence	35.4	27.4
Some Influence	35.8	30.6
Major Influence	28.9	42.0
School District Input On Deciding How		
School Discretionary Funds Will Be Spent		
No Influence	29.0	79.5
Some Influence	38.1	15.6
Major Influence	32.9	4.9
School-Based Management Committee		
Input On Deciding How School Discretionary		
Funds Will Be Spent		
No Influence	25.2	76.3
Some Influence	27.5	9.7
Major Influence	47.3	14.0

Table 6
Percentages of U.S. Public And Private Elementary School Principal Responses Regarding
Professional Development

Decision Makers	Public	Private
Administrator Input On Professional Development		
No Influence	.4	0.0
Some Influence	20.8	7.0
Major Influence	78.8	93.0
Teacher Input On Professional Development		
No Influence	3.6	2.2
Some Influence	29.0	33.9
Major Influence	67.5	64.0
Parent Input On Professional Development		
No Influence	68.4	78.3
Some Influence	28.0	19.3
Major Influence	3.7	2.5
School Board Member Input On		
Professional Development		
No Influence	35.1	52.6
Some Influence	42.3	35.3
Major Influence	22.6	12.2
School District Input On		
Professional Development		
No Influence	8.2	43.2
Some Influence	28.7	27.3
Major Influence	63.1	29.5
School-Based Management Committee Input		
On Professional Development		
No Influence	22.8	82.4
Some Influence	28.4	9.9
Major Influence	48.7	7.7

Pearson chi-squares were conducted to ascertain the extent to which differences were present between public and private elementary school principals for each individual decision and each individual decision-maker. This procedure permitted a detailed analysis of where specific differences might be in school-based management implementation.

Six Pearson chi-squares were calculated to determine whether public and private elementary school principals reported a different amount of principal influence (i.e., 0, 1, or 2) in each of the six decision categories. The first chi-square revealed a statistically significant difference between public and private elementary school principals in the degree of principal influence regarding establishing criteria for the hiring and firing of teachers, $\chi^2(2) = 17.23$, $p < .0001$. As reported in Table 1, private elementary school principals (91.8%) indicated they had significantly more influence in the hiring and firing of teachers than was indicated by the public elementary school principals (78.5%). A second chi-square yielded a statistically significant difference in the degree of principal influence regarding the selection of textbooks, $\chi^2(2) = 78.60$, $p < .0001$. As depicted in Table 2, private elementary school principals (83.1%) indicated they had significantly more influence in the selection of textbooks than was indicated by public elementary school principals (46.2%). A third chi-square revealed the presence of a statistically significant difference in the degree of principal influence in the setting of curricular guidelines and standards, $\chi^2(2) = 72.07$, $p < .0001$. As depicted in Table 3, private elementary school principals (88.6%) indicated that they had significantly more influence in the setting of curricular guidelines and standards than was indicated by the public elementary school principals (54.1%).

In the fourth chi-square, a statistically significant difference was noted in the degree of principal influence on establishing policies and practices for student grading and evaluation, $\chi^2(2) = 56.58$, $p < .0001$. As shown in Table 4, private elementary school principals (88.4%) indicated that they had significantly more influence in establishing policies and practices for grading and evaluation than was reported by the public elementary school principals (58.4%).

345

In terms of school discretionary funds, no statistically significant difference was noted between public and private elementary school principals. See Table 5 for exact percentages. Regarding professional development, a chi-square yielded a statistically significant difference in the degree of principal influence between public and private elementary school principals, $\chi^2(2) = 19.42$, $p < .0001$. As depicted in Table 6, private elementary school principals (93.0%) indicated that they had significantly more influence in professional development planning than was indicated by the public elementary school principals (78.8%). The effect sizes for the five statistically significant differences between public and private elementary school principals ranged from small (hiring and firing; policies and practices for grading; professional development) to moderate (selection of textbooks; curricular guidelines and standards) in size (Cohen, 1988).

Another chi-square revealed a statistically significant difference between public and private elementary school principals in the degree of teacher influence regarding establishing criteria for the hiring and firing of teachers, $\chi^2(2) = 25.05$, $p < .0001$. As reported in Table 1, public elementary school principals (23.0%) indicated that teachers had significantly more influence in the hiring and firing of teachers than was indicated by the private elementary school principals (11.6%). A second chi-square yielded a statistically significant difference in the degree of teacher influence regarding the selection of textbooks, $\chi^2(2) = 6.03$, $p < .0001$. As depicted in Table 2, private elementary school principals (83.3%) indicated that they had significantly more influence in the selection of textbooks than was indicated by the public elementary school principals (74.6%). A third chi-square revealed the presence of a statistically significant difference in the degree of teacher influence in the setting of curricular guidelines and standards, $\chi^2(2) = 14.74$, $p < .0001$. As depicted in Table 3, private elementary school principals (69.2%) indicated that they had significantly more influence in the setting of curricular guidelines and standards than was indicated by the public elementary school principals (52.8%). In the fourth chi-square, a statistically significant difference was not noted in the degree of teacher influence on establishing policies and practices for student grading and evaluation. See Table 4, for exact percentages.

In terms of school discretionary funds, a statistically significant difference was noted between public and private elementary school principals regarding teacher influence, $\chi^2(2) = 21.07$, $p < .0001$ as depicted in Table 5. Regarding professional development, a chi-square did not yield a statistically significant difference in the degree of teacher influence between public and private elementary school principals. See Table 6 for exact percentages. The effect sizes for the five statistically significant differences between public and private elementary school teachers were small (hiring and firing; textbooks; curricular guidelines and standards; discretionary funds, and professional development) were small in size (Cohen, 1988).

Another chi-square did not reveal a statistically significant difference between public and private elementary school principals in the degree of parent influence regarding establishing criteria for the hiring and firing of teachers. See Table 1 for percentages. A second chi-square yielded a statistically significant difference in the degree of parent influence regarding the selection of textbooks, $\chi^2(2) = 7.56$, $p < .0001$. As depicted in Table 2, public elementary school principals (6.3%) indicated that they had significantly more influence in the selection of textbooks than was indicated by private elementary school principals (4.3%). A third chi-square revealed the presence of a statistically significant difference in degree of parent influence in the setting of curricular guidelines and standards, $\chi^2(2) = 20.66$, $p < .0001$. As depicted in Table 3, public elementary school principals (7.3%) indicated that they had significantly more influence in the setting of curricular guidelines and standards than was indicated by the private elementary school principals (2.4%).

In the fourth chi-square, a statistically significant difference was noted in the degree of parent influence on establishing policies and practices for student grading and evaluation, $\chi^2(2) = 28.35$, $p < .0001$. As shown in Table 4, public elementary school principals (7.3%) indicated that they had significantly more influence in establishing policies and practices for grading and evaluation than was reported by the private elementary school principals (3.1%).

In terms of school discretionary funds, no statistically significant difference was noted between public and private elementary school principals. See Table 5 for exact percentages. Regarding professional development, no statistically significant difference was noted between public and private elementary school principals. See Table 6 for exact percentages. The effect sizes for the three statistically significant differences between public and private elementary school principals (selection of textbooks; curricular guidelines and standards; and policies and practices for grading) were small in size (Cohen, 1988).

Another first chi-square revealed a statistically significant difference between public and private elementary school principals in the degree of school board influence regarding establishing criteria for the hiring and firing of teachers, $\chi^2(2) = 52.73$, $p < .0001$. As reported in Table 1, public elementary school principals (60.9%) indicated that they had significantly more influence in the hiring and firing of teachers than was indicated by the private elementary school parents (33.1%). A second chi-square yielded a statistically significant difference in the degree of school board influence regarding the selection of textbooks, $\chi^2(2) = 71.49$, $p < .0001$. As depicted in Table 2, public elementary school principals (38.6%) indicated that they had significantly more influence in the selection of textbooks than was indicated by the private elementary school principals (9.6%). A third chi-square revealed the presence of a statistically significant difference in the degree of school board influence in the setting of curricular guidelines and standards, χ^2 (2) = 105.02, $p < .0001$. As depicted in Table 3, public elementary school principals (62.7%) indicated that they had

significantly more influence in the setting of curricular guidelines and standards than was indicated by the private elementary school principals (21.0%).

In the fourth chi-square, a statistically significant difference was noted between public and private elementary school principals regarding school board influence on establishing policies on student grading, $\chi^2(2) = 121.93$, $p < .0001$. As depicted in Table 4, public elementary school principals (57.7%) indicated that they had significantly more influence on establishing policies on student grading than was indicated by the private elementary school principals (18.1%).

In terms of school discretionary funds, a statistically significant difference was not noted between public and private elementary school principals. See Table 5 for percentages. Regarding professional development, a statistically significant difference was noted between public and private elementary school principals regarding school board influence on professional development, $\chi^2(2) = 17.12$, $p < .0001$. As depicted in Table 6, public elementary school principals (22.6%) indicated that they had significantly more influence on establishing policies on student grading than was indicated by the private elementary school principals (12.2%). The effect sizes for the five statistically significant differences between public and private elementary school principals were small (hiring and firing teachers; and professional development) were small in size. The effect sizes for selection of textbooks; curricular guidelines and standards; and policies for student grading were moderate (Cohen, 1988).

Another chi-square revealed a statistically significant difference between public and private elementary school principals in the degree of school district influence regarding establishing criteria for the hiring and firing of teachers, $\chi^2(2) = 135.58$, $p < .0001$. As reported in Table 1, public elementary school principals (67.8%) indicated that they had significantly more influence in the hiring and firing of teachers than was indicated by the private elementary school principals (25.4%). A second chi-square yielded a statistically significant difference in the degree of school district influence regarding the selection of textbooks, $\chi^2(2) = 115.93$, $p < .0001$. As depicted in Table 2, public elementary school principals (57.8%) indicated that they had significantly more influence in the selection of textbooks than was indicated by the private elementary school principals (18.4%). A third chi-square revealed the presence of a statistically significant difference in the degree of school district influence in the setting of curricular guidelines and standards, $\chi^2(2) = 139.94$, $p < .0001$. As depicted in Table 3, public elementary school principals (80.0%) indicated they had significantly more influence in the setting of curricular guidelines and standards than was indicated by private elementary school principals (42.1%).

In the fourth chi-square, a statistically significant difference was noted between public and private elementary school principals regarding school district influence on establishing policies on student grading, $\chi^2(2) = 138.73$, $p < .0001$. As depicted in Table 4, public elementary school principals (71.5%) indicated that they had significantly more influence on establishing policies on student grading than was indicated by the private elementary school principals (37.9%).

In terms of school discretionary funds, a statistically significant difference was noted between public and private elementary school principals regarding the spending of school discretionary funds, $\chi^2(2) = 107.48$, $p < .0001$. As depicted in Table 5, public elementary school principals (32.9%) indicated that they had significantly more influence on spending school discretionary funds than was indicated by the private elementary school principals (4.9%). Regarding professional development, a statistically significant difference was noted between public and private elementary school principals regarding school district influence on professional development, $\chi^2(2) = 103.68$, $p < .0001$. As depicted in Table 6, public elementary school principals (63.1%) indicated that they had significantly more influence on professional development than was indicated by private elementary school principals (29.5%). The effect sizes for the six statistically significant differences between public and private elementary school principals were moderate (hiring and firing teachers; selection of textbooks; curricular guidelines and standards; policies for student grading; discretionary school funds; and professional development) were moderate in size (Cohen, 1988).

Another chi-square revealed a statistically significant difference between public and private elementary school principals in the degree of school-based management committee influence regarding establishing criteria for the hiring and firing of teachers, $\chi^2(2) = 17.95$, $p < .0001$. As reported in Table 1, public elementary school principals (15.9%) indicated that they had significantly more influence in the hiring and firing of teachers than was indicated by the private elementary school principals (7.5%). A second chi-square yielded a statistically significant difference in the degree of school-based management committee influence regarding the selection of textbooks, $\chi^2(2) = 55.28$, $p < .0001$. As depicted in Table 2, public elementary school principals (33.5%) indicated that they had significantly more influence in the selection of textbooks than was indicated by the private elementary school principals (4.3%). A third chi-square revealed the presence of a statistically significant difference in the degree of school-based management committee influence in the setting of curricular guidelines and standards, $\chi^2(2) = 62.13$, $p < .0001$. As depicted in Table 3, public elementary school principals (30.1%) indicated that they had significantly more influence in the setting of curricular guidelines and standards than was indicated by the private elementary school principals (11.7%).

In the fourth chi-square, a statistically significant difference was noted between public and private elementary school principals regarding the influence of the school-based management committee influence on establishing policies on student grading, $\chi^2(2) = 62.11$, $p < .0001$. As depicted in Table 4, public elementary school principals (30.1%) indicated that they had significantly more influence on establishing policies on student grading than was indicated by

347

the private elementary school principals (8.5%).

In terms of school discretionary funds, a statistically significant difference was noted between public and private elementary school principals regarding school-based management committee influence on the spending of school discretionary funds, $\chi^2(2)$ =88.66, $p < .0001$. As depicted in Table 5, public elementary school principals (47.3%) indicated that they had significantly more influence on spending school discretionary funds than was indicated by private elementary school principals (14.0%). Regarding professional development, a statistically significant difference was noted between public and private elementary school principals regarding school-based management committee influence on professional development, $\chi^2(2) = 120.76$, $p < .0001$. As depicted in Table 6, public elementary school principals (48.7%) indicated that they had significantly more influence on professional development than was indicated by the private elementary school principals (7.7%). The effect sizes for the five statistically significant differences between public and private elementary school principals were moderate (selection of textbooks; curricular guidelines and standards; policies for student grading; discretionary school funds; and professional development) were moderate in size. There was one statistically significant difference between public and private elementary school principals that was small, hiring and firing of teachers (Cohen, 1988).

In sum, differences were present regarding the implementation of school-based management across the United States in public and private elementary schools. Furthermore, differences regarding the influence different decision-makers have in the six areas of decisions made in elementary schools were also reported by all respondents to be present.

Discussion

Public school principals also reported a high degree of involvement by the school-based management committee regarding influence across all six decision categories namely, hiring/firing teachers, selecting textbooks, setting curricular guidelines/standards, establishing policies and practices for student grading/evaluation, deciding how school discretionary funds will be spent, and planning professional development. Responses from private school principals indicated a low degree of school-based management decision-making committee involvement regarding influence across all six-decision categories. This finding may again be due to the lack of federal and state mandates for the implementation of shared governance (Rodriguez, 2000). Revealed within the literature was that federal legislation, state regulations, district mandates, local and community interests, all have demanded change in public schools but not private schools. In addition, because construction of campus improvement plans call for the expertise of many people in a variety of areas, public school principals may be more open to the input of others who are knowledgeable. Private school principals are not required to comply with state regulations (Rodriguez, 2000). After all, school-based management is a structure and process that allows for greater decision-making power related to the areas of instruction, budget, policies, rules and regulation, staffing, and all matters of governance (Herman & Herman, 1994). The more administrators deem stakeholder input to be important, the more likely they may be to empower those stakeholders (Glickman, 1993; Herman & Herman, 1994; Schlechty, 1997; Sergiovanni, 1992 & 1994).

Though different perceptions exist about the role and function of school-based management in schools, no standard operating model exists of shared governance for public or private schools. Murphy and Beck (1995) argued that the elusiveness of decentralized participation as a construct also creates challenges for the SBM implementation process. As a result, this change process involves controversies, conflicts, frustrations, and ultimately satisfaction when educators exert a collective will to do more for all their students (Glickman, 1993). In addition, this reform movement of SBM in the United States is based on the shared belief that the best education grows out of the wisdom, care, and diligence of members of local schools and local communities who take on greater authority, autonomy, and public responsibility for their students (Glickman, 1993). Public school respondents suggest that public schools are generally willing to explore and make changes in their school, whereby private schools are reluctant to change their school environment.

Consequently, some public elementary school principals may also be responding more than private elementary school principals to the low morale of school employees and the decrease in organizational effectiveness, and thus are making school structural changes. Educational systems in the United States have been publicly criticized for being disorganized and having little apathy for the plight of their employees (Conley, 1993; Schlechty, 1997). Therefore, it seems natural that some school principals would consider school-based management as opposed to traditional school structures. Although the implementation of SBM varies from school to school, its focus on collaboration and shared governance are seen as essential to school restructuring (Schlechty, 1997).

Glickman (1993) reported that schools need to make their own judgments regarding the best way to proceed at any particular moment and each school must choose their own model for shared governance. Furthermore school-based decision-making training for committee members often encompasses the construction of improvement plans (Rodriguez, 2000). Though both public and private elementary school principals perceived degrees of involvement by committees, public school respondents indicated a higher degree of school-based management implementation than private school respondents. The difference may be that some public elementary schools foster educational citizenry for a democracy and attempt to model the concept of shared governance in their school whereas others do not value democracy as a priority belief (Conley, 1993). Private elementary schools may choose to maintain a neutral position and stay true to the philosophy for their school. A democratic form of school governance strives for decisions that focus on matters of school-wide education, is fair and equal in the distribution of power and is morally consistent with the goal of democratic engagement of students (Glickman, 1993).

Percentages of public and private elementary school principal responses regarding the influence of decision-makers on the hiring/firing of teachers, selection of textbooks, setting curricular guidelines and standards, establishing policies and practices for student grading/evaluation, deciding how school discretionary funds will be spent, and planning for professional development were investigated. Public elementary school principals indicated a higher degree of influence by teachers, school boards, school districts, and school-based management committees on the hiring/firing of teachers at their school. Private elementary school principals reported a higher degree of influence for the principal and parents. In contrast, in a study by the National Center for Education Statistics (1997) when comparing ratings for 1987-88 to those ratings for 1993-94, evidence was present of an increase in public school principal influence over hiring new teachers (5.3 versus 4.9). Perhaps this difference in perception may be due to the fact that private school principals are expected to be the individual responsible for their school teaching staff and only need to respond to the parents who pay tuition for their children's education. Public elementary school principals view themselves and their school as only one voice in many with regard to the hiring/firing of teachers at their school (Sergiovanni, 1994). In contrast, Synder and colleagues (1997) reported a different view on the issue of hiring/firing of teachers when investigating the condition of public and private schools in 1997. Only a small percentage of public and private elementary school teachers were likely to think that they had a great deal of influence over the hiring/firing of teachers.

Public elementary school principals also indicated a higher degree of influence by parents, school boards, school districts, and school-based management committees on the selection of textbooks. Private elementary school principals reported a higher degree of influence for the principal and teachers. This influence may suggest that public elementary schools have progressed to a level of partnership with their school district personnel and school board in regard to shared governance, whereby private schools are not motivated to include various stakeholders in their decision-making (Sergiovanni, 1994). Perhaps this difference in perception may be explained by the fact that private elementary school principals are expected to be the person responsible to a lesser degree along with their teachers (Snyder, 1997). Additionally, private elementary schools do not have state mandates on the selection of their textbooks. But public elementary schools must comply with the use of state-selected books (Baker et al., 1996). Textbook companies are big business in public school education. In contrast, research by Synder and colleagues (1997) indicated different findings concerning textbook selection; they discovered that relatively few teachers in public and private schools thought that they had a good deal of control over the selection of textbooks.

Public elementary school principals again indicated a higher degree of influence by parents, school boards, school districts, and school-based management committees on setting curricular guidelines and standards. Private elementary school principals again reported a higher degree of influence for the principal and teachers. Snyder and colleagues (1997), using national data survey results, agreed that private school principals were more likely to report that they, rather than any other group, had a great deal of influence on establishing curriculum. In addition, public school principals attributed more influence to the State Department of Education, school district staff (which private schools do not have), and even to teachers than to themselves (Synder, 1997). Therefore, the possibility may exist that the difference in perception may be because private elementary school principals consider themselves to be the sole decision-maker concerning curriculum planning and instruction. Public elementary schools, conversely, are expected to include all stakeholders in the district and the school board to design the school curriculum and instruction.

Public elementary school principals also indicated a higher degree of influence by parents, school boards, school districts, and school-based management committees on establishing policies and practices for student grading/evaluation. Private elementary school principals again reported a higher degree of influence for the principal and teachers. In the 1993-94 national study (NCES) by Synder and colleagues (1997), private school principals and teachers reported that they believed they had a great deal of influence on a number of school policy areas. One can deduce again that the difference in perception may be because private elementary school principals are viewed to be the only decision-makers along with teachers on the establishment of policies and practices for student grading/evaluation by tuition paying parents. Public elementary schools, conversely, are expected to include all stakeholders of the district and the school board to design the establishment of policies and practices for student grading/evaluation (Glickman, 1993; Herman & Herman, 1994; Rodriguez, 2000).

Public elementary school principals once more indicated a higher degree of influence by teachers, school districts, and school-based management committees on deciding how school discretionary funds will be spent. Private elementary school principals in contrast reported a higher degree of influence for the principal, parents, and school board. Therefore, the possibility may exist that the difference in perception may be that private elementary school principals are responsible for the design of the school budget along with parents and the school board, which are usually composed of tuition-paying parents (Baker et al., 1996). Public elementary schools, conversely, are mandated by the state to include all stakeholders of the district and the school board to design the school budget. Public elementary school principals, district personnel and board members are all together accountable to the state and tax-payers for responsible spending of public funds (Rodriguez, 2000). In contrast, Synder and colleagues (1997) reported a different view on the issue of fiscal spending when investigating the condition of public and private schools in 1997. Only a small percentage of public and private school teachers were likely to think that they had a great deal of influence over budget spending. Accordingly, some public and private school districts may say they are all in favor of school-based management, as long as they do not have to do anything differently. This unwillingness to look at underlying assumptions, values, beliefs, practices, and relationships can prevent schools from coming to grips with the profound and disturbing implications of true restructuring (Conley, 1993).

Public elementary school principals yet again indicated a higher degree of influence by teachers, parents, school

boards, school districts, and school-based management committees on planning professional development. Private elementary schools in contrast reported a higher degree of influence for only the principal. It is possible that private elementary school principals again view themselves as the sole person responsible for the planning of professional development for their teachers as parents hold them accountable for the instructional program at their school. Public schools, conversely, are expected to include all stakeholders with the assistance of district personnel and the school board to plan the professional development of teachers at their campus (Conley, 1993). Findings from the 1997 national report of public and private schools indicated that on certain measures, public school teachers appear to be more qualified in terms of their education than their private school counterparts. Accordingly, public school teachers were also more likely to participate in professional development activities. They believe that teachers, as professionals, should update and improve their teaching skills throughout their career (Sergiovanni, 1994). Snyder and colleagues continued to report that beginning teachers in public schools (those teachers in their first 3 years of teaching) were much more likely than their private school counterparts to participate in a formal teacher induction program (56% versus 29%). However, induction may be done informally in some schools. A possible explanation for public school teacher participation in professional development could be that teachers have a sense of ownership in their professional development and private school teachers do not experience this ownership for their professional training.

Public schools are in the process of second-order change or restructuring. This change might explain a high degree of implementation of school-based management by public schools. They are altering the ways in which schools are put together, including the development of new goals, structures, and roles as opposed to the first-order change, which may be found in private schools with a traditional form of school governance. Conley (1993) reported that first-order change improves the efficiency and effectiveness of what is already occurring without disturbing the basic organizational features, without substantially altering the way that children and adults perform their roles.

School-based management is commonly applied to only a small subset of the constellation of decisions that go into running a school (Bimber, 1993). Consequently, some school districts have decentralized budgetary decisions but not decisions about personnel or curriculum. Some have decentralized aspects of curriculum only, and others have decentralized other different combinations. Bimber (1993) argued that often SBM plans give authority to schools over marginal issues only; for example safety, and career education. Accordingly, shared decision-making generally does little to change the fact that most schools have discretion over much less than 10% of the money spent within their walls (Bimber, 1993).

Implications for Future Research

Value exists in employing multiple methods and multiple perspectives to produce a more focused and realistic understanding of issues challenging education. Miles and Huberman (1994) are strong advocates of the developmental mixed methods design, where researchers incorporate alternating quantitative and qualitative phases, which build on, and inform one another to produce superior results. The design of this quantitative study evolved from prior quantitative research; the findings now allude to questions that could be answered through one-to-one interviews with a purposefully selected sample of principals, teachers, and parents (Miles & Huberman, 1994).

For example, the administrator response to the *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99* questionnaire suggests topics for further research. The fact that the administrator of both public and private schools indicated low degrees of parent involvement in their school-based management committees raises questions concerning the inclusion of all stakeholders. Why are parents not more involved in the decision-making process at their school?

In addition, the differences discovered in this study regarding the decision-making influence of various stakeholders in school-based management committees requires some further investigation to understand better the environment of our nation's schools. A need exists to explore the training of school-based management committee members. Training for SBM committee members and school staff serving on decision-making committees appears not to be prevalent based on a review of the literature (Conley, 1993).

An additional area for future research is to explore the effects of school-based management on student performance using qualitative research at individual schools in the United States and at different stages of school-based implementation (Rodriguez, 2000). The majority of research on shared governance has focused on process and not product. Rodriguez (2000) suggested that the literature on this topic illustrates a profusion of material on what should occur, how to do it, and the practices that effective school-based managed schools should engage in. The question to ask is whether or not students who attend public and private schools that implement the school-based management model receive a better education than students who attend schools that follow the traditional model.

The theory behind school-based management implies that school leadership is the key to implementation of shared governance in our elementary public and private schools (Conley, 1993; Deal & Peterson, 1994; Herman & Herman, 1994). Shared leadership should be an important research focus (Sergiovanni, 1994). Researchers could empirically be examined by researchers regarding the concept of shared governance and its contribution to school climate, school development, and school effectiveness at the elementary level. Furthermore, a close investigation of the relationship between the school leadership role and the model of school effectiveness at not only the national level but also at the state level could aid in the improvement of effective leadership. For example, Rodriguez (2000) examined shared

governance in the state of Texas as reported in this study. Accordingly, the next release of data from the *Early Childhood Longitudinal Study: Kindergarten 1999-2000* from the National Center of Education Statistics will provide another opportunity to obtain a portrait of shared governance in public and private schools in the United States for that period of time and to examine changes in the principalship since 1998-1999.

Finally, another area for research relates to the educational reform initiatives and shared decision-making. That is, researchers could focus on specific reform initiatives and investigate the extent to which shared decision-making changes or has changed as a result of the reform initiative. It may be that states or schools actively involved in educational reform may have greater shared decision-making practices than those states or schools not as involved in educational reform.

Conclusions

According to our findings, public school principals have implemented school-based management to a higher degree than private schools. Furthermore, survey responses from public school principals, as a whole, indicated a higher degree of implementation regarding the influence and involvement of decision-makers on the six categories of decisions made at their schools: hiring/firing of teachers, selection of textbooks, setting curricular guidelines/standards, establishing policies and practices for student grading/evaluation, deciding how school discretionary funds will be spent, and professional development planning.

From this study, new insights regarding the extent to which principals implement school-based management and the inclusion of stakeholders in school-based management committees across the United States were established. These new insights provide an authentic context from which to conduct further study of school-based management in our public and private schools on the state and national levels.

References

Baker, D., Han, M., & Keil, C. T. (1996). U.S. Department of Education, National Center for Education Statistics, *How different, how similar? Comparing key organizational qualities of American public and private secondary schools*, Washington, D.C.

Barth, R. S. (1990). *Improving schools from within: Teachers, parents, and principals can make the difference*. San Francisco: Jossey-Bass Publishers.

Bimber, B. (1993). *School decentralization: Lessons from the study of bureaucracy*. Santa Monica, California: RAND Institute.

Blasé, J., & Blasé, J. (1997). *The fire is back! Principals sharing school governance*. Thousand Oaks, California: Corwin Press.

Brown, D. J. (1990). *Decentralization and school-based management*. London: The Falmer Press.

Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good*. Mass.: Harvard University Press.

Carlson, R. V. (1996). *Reframing & reform: Perspectives on organization, leadership, and school change*. New York: Longman Publishers.

Chapman, J. (1990). School-based decision-making and management: Implications for school personnel. In C. Chapman (Ed.), *School-based decision-making and management*. London: Falmer.

Cheng, Y. C. (1996). *School effectiveness and school based management: A mechanism for development*. New York: The Falmer Press.

Choy, S. P. (1998). Public vs. private schools. *Principal, 77* (5), 1-4.

Clark, D. L., & Meloy, J. M. (1989). Renouncing bureaucracy: A democratic structure for leadership in schools. In T. J. Sergiovanni, & J. A. Moore, (Eds.) *Schooling for tomorrow: Directing reform to issues that count*. Boston: Allyn & Bacon.

Clune, H. H., & White, P. A. (1988). *School-based management: Institutional variations, implementation, and issues for further research*. New Brunswick, NJ: Center for Policy Research in Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* New York: John Wiley.

Conley, D. T. (1993). *Roadmap to restructuring: Policies, practices and the emerging visions of schooling.* Eugene, Oregon: ERIC Clearinghouse on Educational Management.

Dade County Public Schools (1989). *Renaissance in education.* Miami, Florida: Author.

Deal, T. E., & Peterson, K. D. (1994). *The leadership paradox: Balancing logic and artistry in schools.* San Francisco, CA: Jossey-Bass Publishers.

Fernandez, J. A. (1989, May). Dade County public schools' blueprint for restructured schools. Paper presented at the conference on choice and control in American education, University of Wisconsin-Madison. In J. Murphy, & L. Beck, *School-based management as school reform.* Thousand Oaks, CA: Corwin Press, Inc.

Glickman, C. D. (1993). *Renewing America's schools: A guide for school-based action.* San Francisco: Jossey-Bass Publishers.

Guthrie, J. W. (1986, December). School-based management: The next needed education reform. *Phi Delta Kappan, 68* (4), 305-309.

Hammer, C. (1997). U.S. Department of Education, National Center for Education Statistics, *Public and private school principals in the United States: A statistical profile, 1987-88 to 1993-94,* Washington, D.C.

Hannaway, J. (1992, March). *Decentralization in education: Technical demands as a critical ingredient.* (ERIC Document Reproduction Service No. ED 345 362).

Herman, J. J., & Herman, J. L. (1994). *Making change happen: Practical planning for school leaders.* Thousand Oaks: Corwin Press, Inc.

Hess, G. A. Jr. (1991, February). *School restructuring, Chicago style: A midway report.* Chicago: Chicago Panel on Public School Policy and Finance.

Hill, P. T., & Bonan, J. (1991). *Decentralization and accountability in public education.* Santa Monica, CA: RAND.

Lindelow, J. (1981). School-based management. In Smith, S. C., Mazzarella, J. A., & Piele, P. K. (Eds.) *School leadership: Handbook for survival.* Eugene: University of Oregon, ERIC Clearinghouse on Educational management.

Livingston, M. J., Slate, J. R., & Gibbs, A. (1999). Shared decision-making and practices of rural school principals. *Rural Educator, 21,* (1), 20-26.

Malen, B., Ogawa, R. T., & Kranz, J. (1989). *What do we know about school-based management? A case study of the literature- A call for research.* Paper presented at the conference on Choice and Control in American Education, University of Wisconsin-Madison, May 17-19.

Miles, M. B., & Huberman, M. (1994). *Qualitative data analysis,* Thousand Oakes, California: Sage Publications, Inc.

Murphy, J. (1991). *Restructuring schools: Capturing and assessing the phenomena.* New York: Teachers College Press.

Murphy, J., & Beck, L. G. (1995). *School-based management as school reform: Taking stock.* Thousand Oaks, CA: Corwin Press.

Reynolds, L. J. (1997). *Successful site-based management: A practical guide.* Thousand Oaks: Corwin Press, Inc.

Rodriguez, T. A. (2000). *The implementation of site-based management across Texas: An empirical study* (doctoral dissertation , University of Texas at El Paso, El Paso, TX).

Schlechty, P. C. (1997). *Inventing better schools: An action plan for educational reform.* San Francisco: Jossey-Bass Publishers.

Senge, P. (1990). *The fifth discipline: The art and practice of the learning organization.* New York: Doubleday.

Sergiovanni, T. J. (1992). *Moral leadership: Getting to the heart of school improvement.* San Francisco: Jossey-Bass Publishers.

Sergiovanni, T. J. (1994). *Building community in schools.* San Francisco: Jossey-Bass Publishers.

Sergiovanni, T. J. (2001). *The principalship.* Needham Heights, MA: Allyn and Bacon Publishers.

Smith, W. E. (1993, April). *Teachers' perceptions of role change through shared decision making: A two-year case study.* A paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia.

Snyder, T. (1997). U.S. Department of Education, National Center for Education Statistics, *Findings from the condition of education 1997: Public and private schools: How do they differ?*, Washington, D.C.

Steffy, B. E. (1993). *The Kentucky education reform: Lessons for America.* Lancaster, PA: Technomic.

Texas Education Code. (1991). Chapter 21, Section 21.931.

Texas Education Agency. (1992). *Resource guide on site-based decision-making and district and campus planning.* Austin, Texas (II-1).

Texas Education Agency. (1997-1998). *Snapshot view of Texas.* Austin, Texas.

U.S. Department of Education. (1998, May). *Turning around low-performing schools: A guide for state and local leaders.* Washington, D.C.: Planning and Evaluation Service.

Vann, A. S. (1999, Fall). Shared decision-making: A paper tiger? *Educational Horizons, 78* (1), 16-17.

Watkins, J. M., & Lusi, S. F. (1989, May). *Facing the essential tensions: Restructuring from where you are.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

West, J., Denton, K., & Germino-Hausken, E. (2000, April). U.S. Department of Education, National Center for Education Statistics. *America's Kindergartners: Findings from the Early Childhood Longitudinal Study: Kindergarten Class of 1998-1999*, NCES 2000-070 (Revised), Project Officer, Jerry West. Washington, DC.

Wohlstetter, P., & Buffet, T. (1991, March). *School-based management in big city districts: Are dollars decentralized too?* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

About the Authors

Mary T. Apodaca-Tucker
New Mexico State University

Mary T. Apodaca-Tucker obtained her Ed.D. in Educational Administration from the University of Texas at El Paso in 2001. She is currently a professor at New Mexico State University in Las Cruces, New Mexico teaching Educational Administration courses focused on the Principalship. She has over 30 years of experience teaching and working as an administrator within the El Paso area and has also worked as an Area Superintendent for the Austin ISD. To date, She authored "Understanding Restructuring in Practice: A Study of Teachers' Perceptions of Organizational Interventions and Core Technology on Student Success," Planning and Change (1997) and "Shared Decision-Making: Beliefs and Practices of Principals at the U.S. Mexico Border," *International Electronic Journal for Leadership in Learning* (2001).

John R. Slate
University of Texas at El Paso

John R. Slate received his Ph.D. from the University of Tennessee in Knoxville in 1984. He is currently a Full Professor in the Department of Educational Leadership and Foundations at The University of Texas at El Paso. His current interests are in examining school reform efforts, specifically through analysis of large-scale databases such as state report cards and the National Assessment of Educational Progress.

BEST COPY AVAILABLE

Departamento de Investigación Educativa-
DIE/CINVESTAV
rkent@gemtel.com.mx   kentr@data.net.mx

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Universidad Federal de Rio Grande do Sul-UFRGS
lucemb@orion.ufrgs.br

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Lake Woebeguaranteed:
Misuse of Test Scores in Massachusetts, Part I

Walt Haney[1]
Boston College

Abstract

Misuse of test results in Massachusetts largely guarantees woes for both students and schools. Analysis of annual test score averages for close to 1000 Massachusetts schools for four years (1998–2001) shows that test score gains in one testing period tend to be followed by losses in the next. School averages are especially volatile in relatively small schools (with less than 150 students tested per grade). One of the reasons why scores fluctuate is that the Massachusetts state test has been developed using norm-referenced test construction procedures so that items which all students tend to answer correctly (or incorrectly) are excluded from operational versions of the test. This article concludes with a summary of other reasons why results from state tests, like that in Massachusetts, ought not be used in isolation to make high-stakes decisions about students or schools.

Lake Wobegon is the mythical town in Minnesota popularized by Garrison Keillor in his National Public Radio program "A Prairie Home Companion." It is the town where "all the children are above average" (and "all the women are strong, and all the men, good-looking"). In the late 1980s, it became apparent that Lake Wobegon had come to schools nationwide. For according to a 1987 report by John Cannell, the vast majority of school districts and all states were scoring above average on nationally normed standardized tests (Cannell, 1987). Since it is logically impossible for all of any population to be above average on a single measure, it was clear that something was amiss, that something about nationally normed standardized tests or their use had been leading to false inferences about the test scores of students in the nation's schools. As a result, people came to refer to inflated test results as the Lake Wobegon Wobegon phenomenon. I do not try here to recap the story of Cannell's work on the Lake Wobegon phenomenon and how independent researchers came to verify the phenomenon. (The story is recounted in chapter 7 of Haney, Madaus & Lyons, 1993, for anyone interested).

Rather, my purpose is to introduce a place considerably east of Lake Wobegon; namely, Lake Woebeguaranteed. In this place, the use of state test results in isolation to make important decisions about schools and students pretty well guarantees woes will follow. For as I will explain, such uses of results from what is essentially a norm-referenced test constitute ill-conceived *mis*uses of test results. Before proceeding to this larger story, I recap how the work reported here evolved.

After reading Kane & Staiger (2001), and Bolon (2001), I undertook an analysis of school average scores on the Massachusetts Comprehensive Assessment System (MCAS) grade 4 mathematics tests for 1998, 1999, 2000 and 2001. After summarizing these previous works, I describe the sources of data used in the present analysis, the means by which data were merged from different sources, the analyses undertaken, and the results. The latter confirm the findings by Kane & Staiger (2001) and Bolon (2001); namely, that changes in school average test scores from one year to the next are unreliable indicators of school quality. Next I discuss three reasons this is so, and why misuse of results of the Massachusetts state test virtually guarantees woes for schools and students.[2]

Background

The works that prompted the analyses reported here were Kane & Staiger (2001), and Bolon (2001). The first of these works focused on state test results in North Carolina. North Carolina has an extensive system of testing students, not just with state "competency" tests in grades 3—11, but also with norm-referenced tests in grades 5 and 8. (CCSO, 1998, pp. 19, 21, 22, 24). Students must pass state competency tests in reading and math to graduate from high school. Schools in North Carolina are publicly rated in terms of student test results. However, the paper by Kane and Staiger (2001) from the National Bureau of Economic Research (http://www.nber.org/papers/w8156) shows how misleading these ratings tend to be.

Kane and Staiger analyzed six years worth of student assessment data from the entire state of North Carolina (for nearly 300,000 students in grades 3 through 5). They showed that, regardless of whether results were analyzed in terms of annual results or year to year changes, the test results are mainly random noise—resulting from the particular samples of students who are in tested grades in particular years, and the vagaries of annual test content and administration—not meaningful indication of school quality.

Kane and Staiger concluded with the following four "lessons":

1. Incentives targeted at schools with test scores at either extreme—rewards for those with very high scores or sanctions for those with very low scores—primarily affect small schools and imply very weak incentives for large schools.
2. Incentive systems establishing separate thresholds for each racial/ethnic subgroup present a disadvantage to racially integrated schools. In fact, they can generate perverse incentives for districts to segregate their students.

3. As a tool for identifying best practice or fastest improvement, annual test scores are generally quite unreliable. There are more efficient ways to pool information across schools and across years to identify those schools that are worth emulating.

4. When evaluating the impact of policies on changes in test scores over time, one must take into account the fluctuations in test scores that are likely to occur naturally. (Kane & Staiger, April 2001).

The second work prompting the analyses reported below is Bolon's "Significance of test-based ratings for metropolitan Boston schools" (2001, in *Education Policy Analysis Archives*, http://epaa.asu.edu/epaa/ v9n42/. Also, see Michelson, 2002; Willson & Kellow, 2002 and Bolon, 2002 for discussion of the orginal Bolon article). In this study Bolon examined 1998, 1999, and 2000 MCAS mathematics scores for 47 academic high schools in 32 metropolitan communities in the greater Boston area (vocational high schools were excluded on the grounds that they have a substantially different mission than academic high schools). Bolon found that school average grade 10 MCAS math scores generally changed little over this interval (+1.3 points from 1998 to 1999; and +5.9 points from 1999 to 2000) relative to the range in school average scores (in 1999, for example, school averages ranged from 203 to 254, on the MCAS scale of 200 to 280.) Bolon does note, however, that according to data released by the Massachusetts Department of Education, between 1998 and 2000 grade 10 MCAS math scores rose substantially more than English or science scores (see Bolon's Table 1-1).

Bolon then examined the extent to which seven school characteristics, plus community income (1989), might be used to predict school average grade 10 MCAS math scores. He found that three variables (percent Asian or Pacific Islander, percent limited English proficiency, and per-capita community income) were the only ones statistically significantly related to school average scores (Table 2-12). Together these three variables accounted for 80% of the variance in school average scores. After excluding schools in Boston (for which separate community income data were not available), Bolon found that "by far the strongest factor in predicting tenth grade MCAS mathematics scores is 'per capita community income (1989).' For the schools outside the City of Boston, this factor alone performed nearly as well as all available factors combined, associating 84 percent of the variance compared with 88 percent when all available factors were used."

The study reported here builds on both of the works just discussed. For example, an analytical approach applied by Kane and Staiger to data from North Carolina, namely comparing school size with changes in annual score averages, is employed here. Additionally, while Bolon examined average MCAS scores for Massachusetts high schools, this inquiry addresses MCAS averages for elementary schools.

There are three broad reasons why elementary school average test scores might be more useful indicators of school quality than test averages for high schools. First is the simple fact that there are more elementary schools than high schools. In his study, Bolon analyzed test scores for less than 50 high schools. In contrast, MCAS scores are available for around 1000 elementary schools in Massachusetts. A larger sample offers greater potential to discern meaningful differences in school quality.

The second reason for hypothesizing that grade 4 test scores may be better indicators of school quality than grade 10 test scores is the extent of institutional experience that they may reflect. Children typically enter school in Massachusetts in kindergarten. This means that by spring of grade 4, they have almost five years of education in a particular elementary school (presuming, of course, they did not switch schools). In contrast, grade 10 test scores typically reflect just two years' experience in high school. So on this count, grade 4 test score averages clearly have more potential to reflect differences in school quality than grade 10 score averages.

The third reason for thinking that grade 4 test scores may be better indicators of school quality than grade 10 test scores is that by grade 10 (roughly age 16) individuals' standardized test scores have become relatively fixed, whereas test scores of young children are relatively malleable. This may be illustrated by reference to Benjamin Bloom's classic (1964) work, *Stability and Change in Human Characteristics*. In this book, Bloom reviewed a wide range of evidence on how a number of human characteristics, including height, weight and test scores, tend to change as people age. He showed, for example, that height in the early childhood years tends to be a moderately good predictor of height at maturity, with correlations between height at ages 6 - 10 years and height at age 18 falling in the range of 0.75 to 0.85 for both males and females. Interestingly, height at ages 11-13 for females and 13-15 for males is a less good predictor of height at maturity. This is, of course, due to variation in the ages at which children experience growth spurts as they go through puberty.

In contrast to the physical characteristic of height, mental abilities of young children as measured by standardized tests show relatively little power to predict mental abilities at maturity. Not until around grade 3 or 4 (or age 8 – 9) do children's test scores become relatively reliable predictors of future performance. To provide one example, reading test scores at age 6 (or grade 1) correlate with reading test scores in grade 8 only about 0.65 (Bloom, 1964, p. 98). As Bloom himself put it, "We may conclude from our results on general achievement, reading comprehension and vocabulary development that by age 9 (grade 3) . . . 50% of the general achievement pattern at age 18 (grade 12) has been developed" (Bloom, 1964, p. 105). The relative malleability of young children's test scores suggests that there may be more potential for grade 4 test scores to be affected by school quality, as compared with high schools' effects on grade 10 test scores.

In sum, while Bolon found that school average scores on the Massachusetts' grade 10 state test (MCAS) were not sound indicators of schools quality, there are several reasons for hypothesizing that school average scores for grade 4 might be better indicators of school quality. To test this possibility, the data and analyses described below were employed.

Data Sources

The data used in this study were drawn from four sources. MCAS results for 1998, 1999 and 2000 were drawn from CD data disks issued by the Massachusetts Department of Education entitled "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of May 1998," "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of May 1999," and "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of Spring 2000." The MCAS results for 2001 were drawn from an Excel file named "MCAS2001pub_g4sch01.xls" downloaded from http://boston.com/mcas/ on November 9, 2001. The files from these four sources contain MCAS results for all schools and districts in Massachusetts for 1998, 1999, 2000 and 2001. From these results, grade 4 MCAS math averages were extracted for all schools in Massachusetts.

Math rather than English Language Arts (ELA) test scores were selected for study for two reasons. First, it is reasonably well-established that schools have more influence on math test scores than on English (or at least reading) test scores (Haney, Madaus & Lyons, 1993). Second, it is apparent that there have been a number of problems in past years in the scaling of MCAS grade 4 ELA scores.

The numbers of records for which MCAS grade 4 average results are available from each of the sources mentioned above are as follows:

Year	No. of Records
1998	1336
1999	1355
2000	1366
2001	1049

The reason for more records in 1999 and 2000 than in 1998 is the creation of a number of new elementary schools (mostly charter schools). The file for 2001 is smaller than those for previous years because it included only school average, but not district average scores. Merging records from these four data files proved more difficult than anticipated. Labels for some variables were changed across the years and names for some schools are reported inconsistently in these four sets of data. Nonetheless after examining pairs of records for 1998, 1999, 2000 and 2001, I was able to create a merged data file of MCAS grade 4 math results (and numbers of students tested) for 1998-2001. A copy of this data file is appended to this article for anyone interested in secondary analysis (see Appendix).

The merged data file of grade 4 MCAS math school averages, after deletion of district averages, contained records for 977 schools. Table 1 shows summary descriptive statistics for this data set. As can be seen, the numbers of fourth graders tested per school in these three years ranged from just 10 to 328. The school average MCAS scores ranged from a low of 206 in 1998 to a high of 263 in 2001. Over the four years of MCAS testing, on average, there were initially slight increases in average MCAS scores—a 1.5 point increase, on average, between 1998 and 1999, and an increase of 0.5 of a point between 1999 and 2000, but then level scores between 2000 and 2001. Changes in score averages for individual schools from year to year ranged from a low of—22 to +18 points. As Bolon found with regard to grade 10 MCAS scores, these school average changes in grade 4 MCAS scores are considerably smaller than the range in school average scores, which varied by 50 points or more in all four years of test administration.

Table 1
Summary Statistics on Grade 4
MCAS Math School Averages, 1998–2001

	1998	1999	2000	2001
Number tested per school				
Minimum	11	10	10	11
Maximum	317	309	320	328
Mean	72.0	72.9	73.2	72.6
Median	63	65	65	64
SD	41.1	41.4	42.3	42.5
Average MCAS score				
Count	977	977	977	977
Minimum	206	208	210	213
Maximum	261	260	260	263
Mean	233.1	234.6	235.1	235.1
Median	233	235	236	235
SD	9.67	9.19	9.19	8.32
Change in MCAS Average	1998 to 1999	1999 to 2000	2000 to 2001	
Minimum		-14	-17	-22
Maximum		25	17	18
Mean		1.5	0.5	-0.04
Median		1.0	1.0	0
SD		5.0	4.6	4.6

Figure 1 shows a scatter plot of how 1999 school averages compared with those from 1998. As can be seen, there is a fairly strong relationship between 1998 and 1999 averages. Schools with higher MCAS grade 4 test score averages in 1998 tended to have higher averages in 1999. The correlation between score averages in 1998 and 1999 was 0.860. The regression relationship between score averages in 1999 and 1998 is:

Gd4MCASAvg99 = 44.03 + 0.81(Gd4MCASAvg98)

For statistically inclined readers, it may be noted that these correlation and regression relationships are statistically significant—that is, extremely unlikely that they might occur by chance.
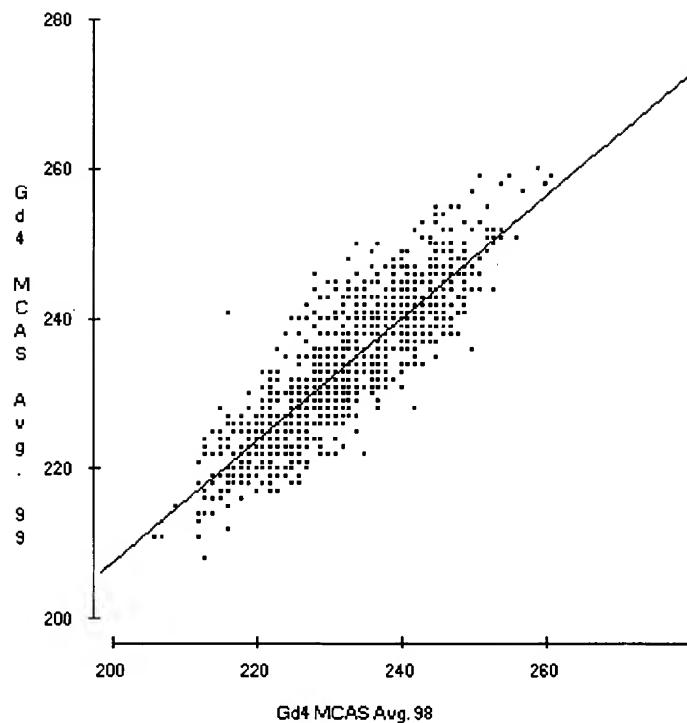
Figure 1. Scatter Plot of School Average Grade 4 MCAS Math Scores 1998 vs. 1999

Figure 2 shows the relationship between grade 4 math MCAS score averages in 1999 and 2000. As can be seen, the relationship between year 2000 MCAS grade 4 math averages and those for 1999 is similar to the 1998-1999 relationship, but even slightly stronger. The correlation between score averages in 2000 and 1999 is 0.875. The regression relationship of average scores in 2000 and 1999 is:

Gd4MCASAvg00 = 29.54 + 0.88(Gd4MCASAvg99)
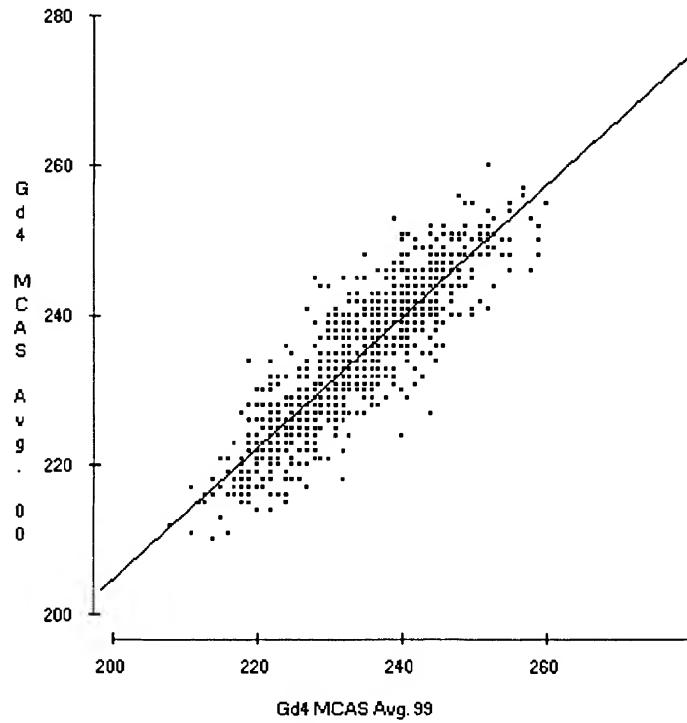
Figure 2. Scatter Plot of School Average Grade 4 MCAS Math Scores 1999 vs. 2000

Figure 3 shows the relationship between score averages in 2001 and 2000. As can be seen, the relationship between score averages in 2001 and 2000 is highly similar to the relationships evident in the previous two pairs of years. The correlation between 2001 and 2000 score averages is. 0. 866. The regression relationship of averages in 2001 and 2000 is:

Gd4MCASAvg01 = 50.8 + 0.78(Gd4MCASAvg00)
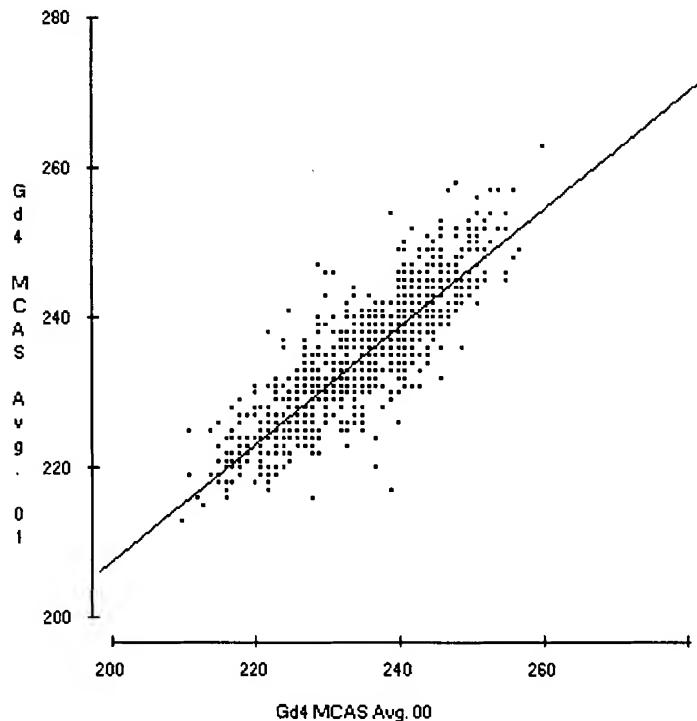
Figure 3. Scatter Plot of School Average Grade 4 MCAS Math Scores 2000 vs. 2001

Next let us consider, á la Kane and Staiger the relationship between school size and change in score averages from one year to the next. For these analyses school size has been calculated simply as the average number of students tested in the two years across which change is calculated.

Figure 4 shows the relationship between change in average MCAS grade 4 scores between 1998 and 1999 and school size (defined as the average of the numbers of students tested in the two years). As can be seen, schools with less than
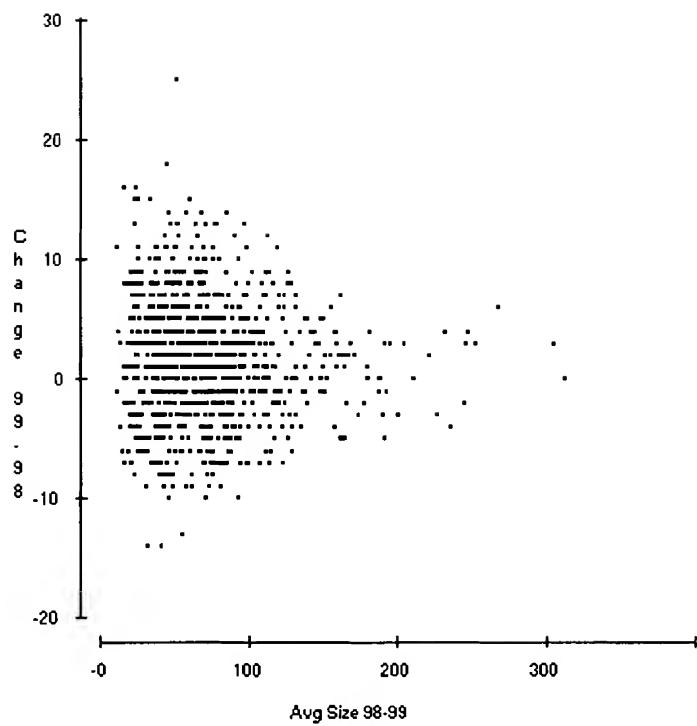
Figure 4. Change in MCAS Grade 4 Math Average Score 1998 to 1999 v. School Size

100 or so students tested show changes in MCAS average scores of as much as 15-20 points. However schools with more than 150 students tested per year show much smaller changes—generally less than 5 points.

Figure 5 shows analogous results for 1999 to 2000 score changes. As can be seen, the pattern shown here is similar to that shown in Figure 4. Schools with smaller numbers of students tested tended to have much more "volatility" (to use Kane and Staiger's phrase) in average scores than schools with larger numbers of students tested.
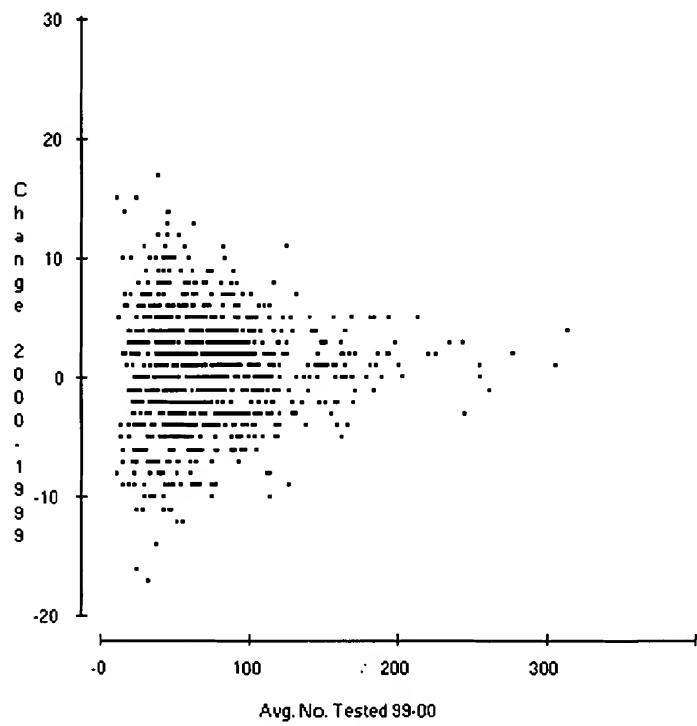
Figure 5. Change in MCAS Grade 4 Math Average Score 1999 to 2000 v. School Size

Figure 6 shows the relationship between school size and change in average grade 4 MCAS scores between 2000 and 2001. The pattern is very similar to that apparent in the previous two figures. Schools with less that 100 students tested showed much larger swings in test score averages than schools with larger numbers of students tested.

Avg. No. Tested 00-01
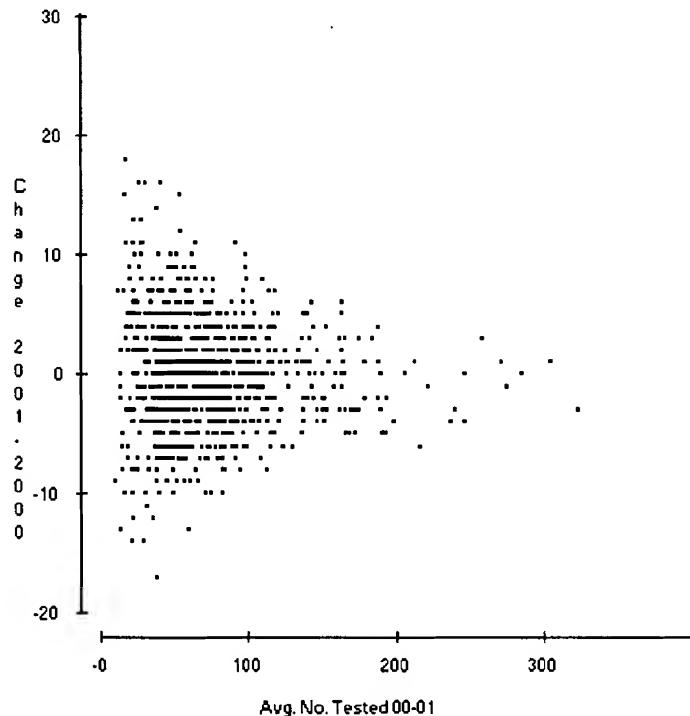
Figure 6. Change in MCAS Grade 4 Math Average Score 2000 to 2001 v. School Size

Given the political prominence of high stakes testing in Massachusetts (as elsewhere), it is not surprising that various observers have tried to use changes in school MCAS scores from one year to the next to identify high quality or "exemplary" schools. For example, in a high profile ceremony at the Massachusetts State House in December 1999, five school principals were presented with gifts of $10,000 each "for helping their students make significant gains on the MCAS" (http://www.doe.mass.edu/news/archive99/Dec99/122299pr.html) (accessed November 15, 2001). Though the cash awards were donated by a private foundation, the ceremony recognizing the five schools was attended by the Massachusetts Governor, Lieutenant Governor and Commissioner of Education. The press release for the event stated: "The schools were recognized as having the highest percentage improvement in overall MCAS scores between 1998 and 1999 in English Language Arts, Mathematics and Science and Technology" (http://www.doe.mass.edu/news/archive99/Dec99/122299pr.html).

Anyone with even a modest knowledge of statistics will note the absurdity of this statement. Since the MCAS scale of 200 to 280 is arbitrary and has no meaningful zero point, it is meaningless to calculate percentage increases in scores. This indicates that whoever in the Massachusetts Department of Education wrote this press release is fundamentally ignorant of statistics—or to be less politically incorrect, in need of improvement in knowledge of statistics. For anyone who has not studied statistics lately and hence may not appreciate the absurdity of calculating percentage increases on arbitrary test score scales, I suggest the following exercise. Calculate the percentage increase in temperature going from 50 degrees Fahrenheit to 68 degrees Fahrenheit. Next, figure out the equivalent temperatures on the Celsius scale and calculate the percentage increase on the Celsius scale. Finally, ask yourself which "percentage increase" is correct.

Four of the five schools receiving the so-called Edgerly awards in 1999 were elementary schools, namely, Riverside Elementary School in Danvers, Franklin D. Roosevelt Elementary School in Boston, Abraham Lincoln Elementary School in Revere, Kensington Elementary School in Springfield. Figure 7 is a recasting of Figure 3, but with these four 1999 Edgerly award schools shown with circles.

363
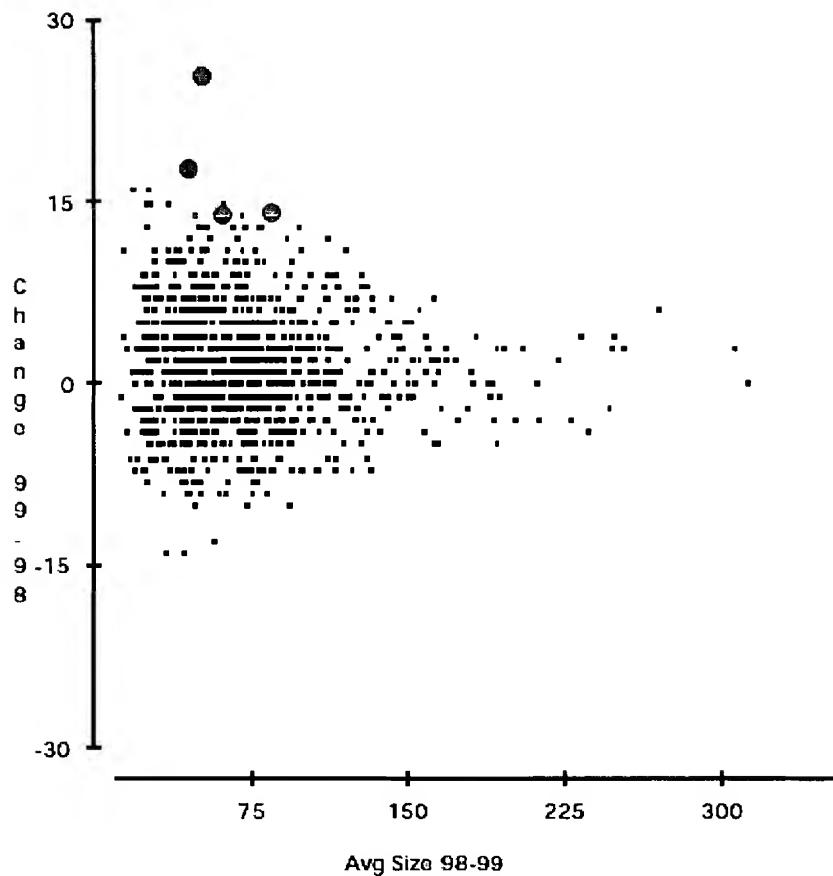
Figure 7. Change in MCAS Grade 4 Math Average Score 1998 to 1999 v. School Size, with Award Schools Marked

As can be seen, the four award schools share two characteristics. First, they are all relatively small schools, each with less than 100 students tested. Second, they showed unusually large score changes from 1998 to 1999. This is not surprising since large MCAS score gains from 1998 to 1999 served as basis for their receiving awards.

Figure 8 recasts Figure 4, again with the 1999 "award" schools marked with circles.
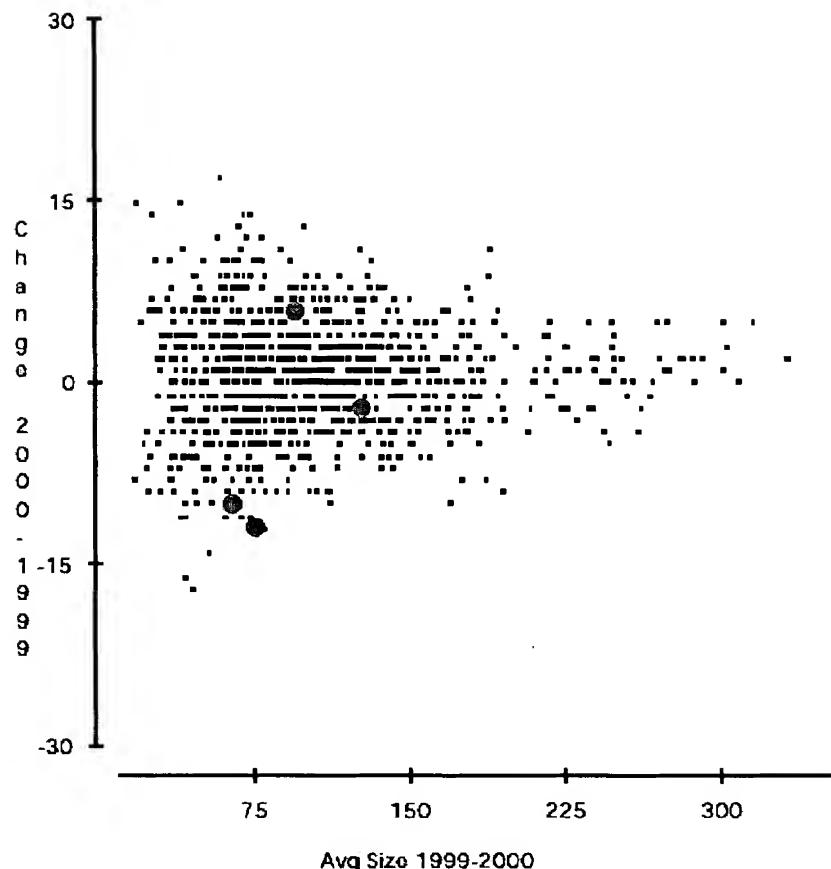
Figure 8. Change in MCAS Grade 4 Math Average Scores 1999 to 2000 v. School Size, with 1999 Award Schools Marked

As can be seen in Figure 8, three out of the four 1999 award schools showed declines in average grade 4 MCAS math scores from 1999 to 2000.

Figure 9 is a variant of Figure 5, showing the relationship between average numbers of students tested in 1999 and 2000 versus the change in average grade 4 math scores between 1999 and 2000. In Figure 9, all of the schools showing a 10 or more point gain in average MCAS scores are marked with circles.
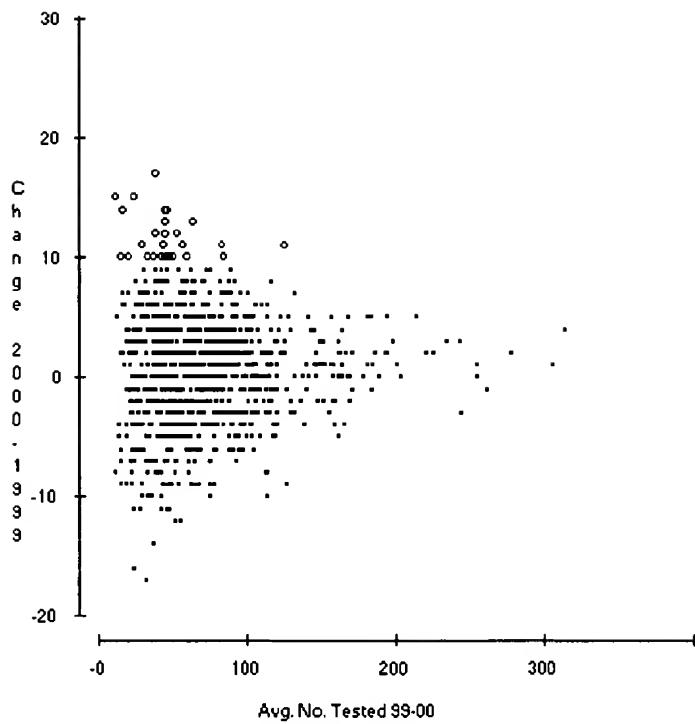
Figure 9. Change in MCAS Grade 4 Math Average Score 1999 to 2000 v. School Size, with Schools showing Gain of 10 Points or More Highlighted with Circles

What happened to these schools the next year? Figure 10 shows change from 2000 to 2001, but with the schools having largest gains from 1999 to 2000 again marked with circles. As can be seen, there were a few schools showing largest gains from 1999 to 2000 that continued to show gains in 2001. But most of the large gain schools from 1999 to 2000, showed declines in 2001. Several of them showed declines from 2000 to 2001 that were just about as large (9-10 points) as were the gains from 1999 to 2000
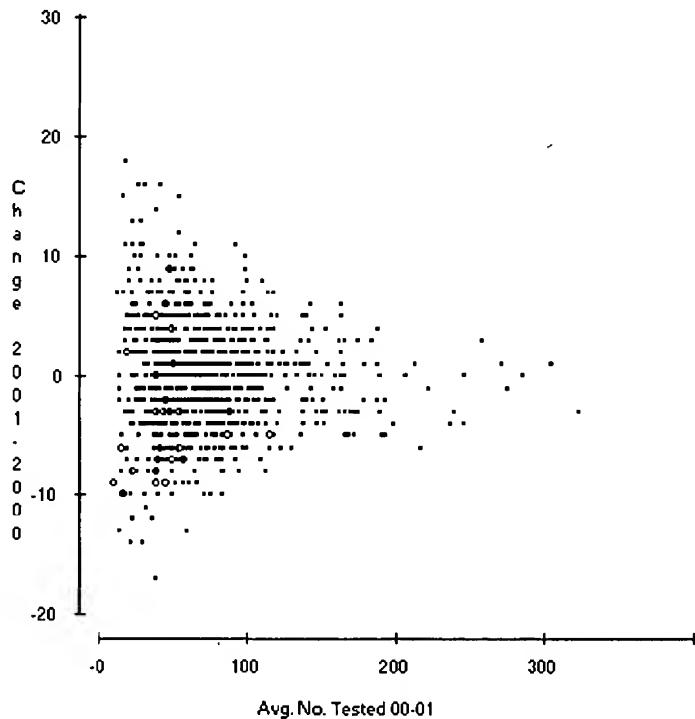
Figure 10. Change in MCAS Grade 4 Math Average Score 2000 to 2001 v. School Size with Schools showing

Gain of 10 Points or More '98 to'99 Highlighted with Circles

Note that almost all of these schools showing large gains in average scores one year, but then large declines the next year, are ones with relatively small numbers of students tested.

The relationship between changes in average scores across pairs of years can be seen more clearly in Figures 11 and 12. Figure 11 shows how change in school average grade 4 MCAS scores between 1998 and 1999 compares with the change between 1999 and 2000. Figure 12 shows how the change between 1999 to 2000 compares with the change between 200 and 2001. As can be seen, there is a negative relationship between change in one interval and changes the next. Schools that show large gains in one interval tend to show losses in the next interval. The correlation between change from 1998 to 1999 and change 1999 to 2000 is -0.388. The correlation for the next pair of years, that is change 1999 to 2000 versus change 2000 to 2001 is -0.396. These negative correlations are both statistically significant.
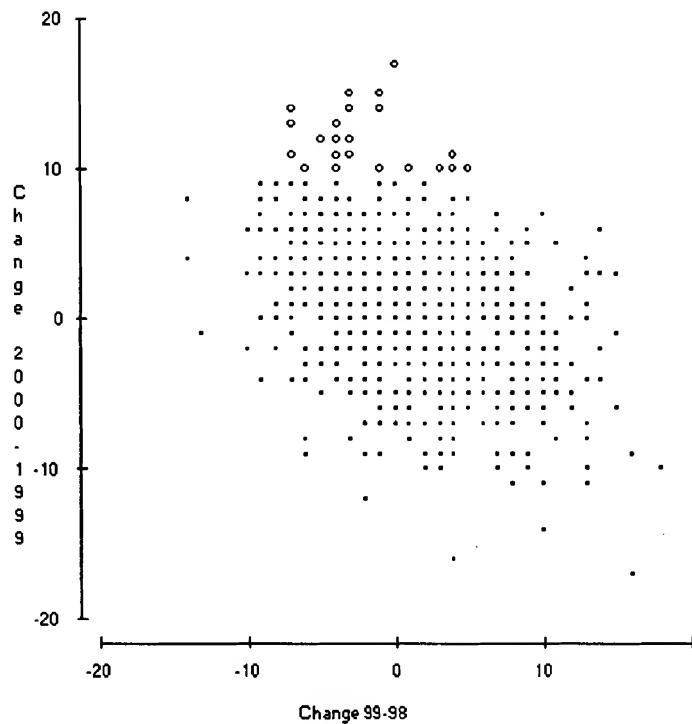
Figure 11. Change in MCAS Grade 4 Math Average Score 1998 to 1999 v. Change 1999 to 2000, with Schools showing Gain of 10 Points or More Highlighted with Circles
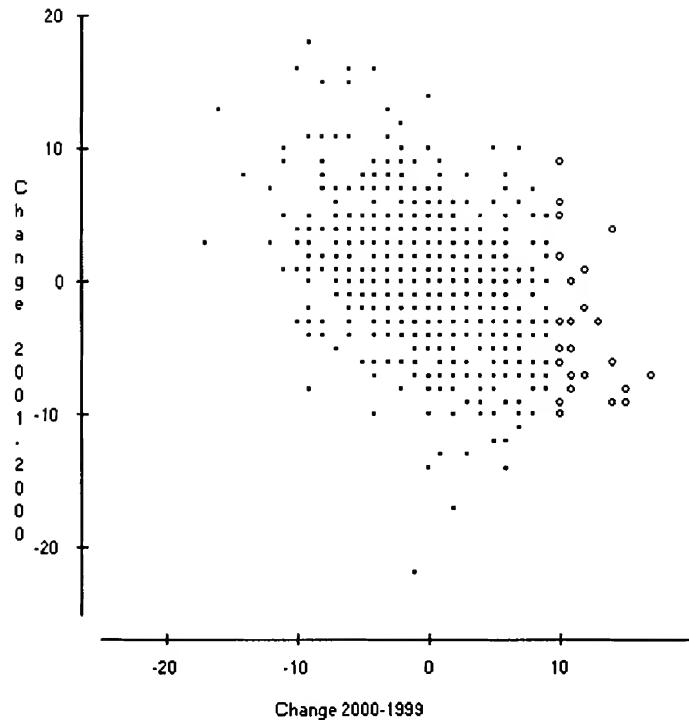
Figure 12. Change in MCAS Grade 4 Math Average Score 1999 to 2000 v. Change 2000 to 2001, with Schools showing Gain of 10 Points or More '98 to '99 Highlighted with Circles

These results are simply a manifestation of the kind of volatility that Kane and Staiger (2001) found in school average test scores in other states. As they found for North Carolina, we have seen above with MCAS scores. School average test scores are particularly volatile for relatively small schools. Moreover schools that show relatively large gains in score averages from one year to the next tend to show losses the following year. sThus, it is clear that school average test scores, or changes in averages from one year to the next, represent poor measures of school quality.

Why are MCAS score averages poor indicators of school quality?

School average test results fluctuate from year to year for several reasons. The most obvious is that one year's class of students will differ from the next. Especially in relatively small schools, with less than 100 students tested per grade, having a few especially test savvy, or not so savvy, students may skew results from one year to the next.

A second likely cause of volatility in school average scores on the Massachusetts test is that the MCAS is of dubious technical merit. When I first examined the 1998 grade 4 English Language Arts (ELA) test, for example, I was surprised to find many poorly worded questions and reading questions for which one did not actually have to read the passage on which they were ostensibly based in order to answer the question (that is, the questions lacked passage dependency). More recently the Massachusetts DOE implicitly acknowledged defects in the 2001 grade 10 ELA and math exams when it dropped one item from each from scoring (http://www.doe.mass.edu/MCAS/01results/threshscore.html). The defective items on the 2001 test were discovered not by the test's developer or state officials but by students (Lindsay, 2001).

More recently, Gallagher (2001) undertook a review of grade 10 MCAS math questions from the 2000 and 2001 test administrations. Gallagher, a professor of Environmental, Coastal and Ocean Sciences, at the University of Massachusetts, Boston, concluded that there were serious problems with 10 to 15% of the grade 10 MCAS math questions. He identified some questions as having wrong answers, some as having more than one correct answer and some as misaligned with the Massachusetts curriculum frameworks. "Overall, my review of these tests indicates that there are serious failures in the choice and review of MCAS questions" (Gallagher, 2001, p. 5, http://www.es.umb.edu/edg/MCAS/mcasproblems.pdf, accessed December 3, 2001).

More generally, the MCAS is not a good indicator of school quality because it has been constructed as a norm-referenced test. Many people assume that the MCAS (and other state-sponsored tests) are criterion referenced tests—that is, tests of well-specified bodies of knowledge and skills. In a paper prepared for an October, 2001 conference at the John F. Kennedy Institute at Harvard University , for example, Kurtz wrote:

> The MCAS, which is known as a "criterion-referenced exam," tests knowledge of a set curriculum and gives students scores based on their level of mastery, in contrast to national "norm-referenced

368

tests," which grade a student's performance in relation to other students. (Kurtz, 2001, p. 6).

However examination of the technical manuals for the MCAS tests, reveals that items have been selected for inclusion on MCAS tests by using norm-referenced test construction procedures. Specifically, items are selected for inclusion on the MCAS in terms of item difficulty and discrimination.

To explain the implications of this—that is, why because of use of norm-referenced test construction procedures, the MCAS is not a "criterion-referenced exam," testing knowledge of a set curriculum and giv[ing] students scores based on their level of mastery," let me discuss at some length the original *1998 Technical Report* on the MCAS (Massachusetts Department of Education, October 1999; hereafter, MDOE, 1999).

The *1998 Technical Report* on the MCAS is a fairly long report, 15 chapters and several appendices. I suspect there are not exactly legions of people who have waded through the whole document. So for those who may not have, let me try to summarize. Before doing so, I note that the *1998 Technical Report* on the MCAS is one of an odd genre of bureaucratic documents that always arouses suspicion, for it bears not a single person's name as author or responsible authority.

The *1998 Technical Report* begins with a "Background and Overview" chapter. Chapter 2 then provides an overview of test design. The first page of this chapter recounts that "The [Massachusetts] Department of Education convened committees of educators from around the state to work with the Department and its testing contractor to design and develop assessments of the learning standards contained in the [Massachusetts] curriculum frameworks" (MDOE, 1999, p. 9). In this chapter it is explained that the MCAS tests were designed to have three different types of items: multiple choice items, short answer items requiring responses from a few words to a few sentences (scored 0 or 1 as incorrect or correct) and open or extended response items requiring responses of up to a half page long (and scored on a 0-4 point scale).

Some 20 pages later (in chapter 6) it is explained that after pilot items were developed, reviewed and tried out, they were screened in terms of a number of statistical characteristics, including item difficulty and discrimination. At this point the *1998 Technical Report* does not make clear exactly what statistical criteria were used in selecting items.

But another 50 or so pages later, in chapter 13, "Item analyses," it becomes considerably clearer how items were selected in terms of difficulty and discrimination. Before offering my own interpretation, let me quote at some length from this chapter.

Difficulty Indices

All multiple-choice, short-answer, and open-response questions were evaluated in terms of difficulty and relationship to overall score according to standard classical test theory practice. Difficulty was measured by averaging the proportion of points received across all students who received the question. Multiple-choice and short-answer questions were scored dichotomously (correct v. incorrect), so for these questions, the difficulty index is simply the proportion of students who correctly answered the question. Open-response questions allowed for scores between 0 and 4. By computing the difficulty index as the average proportion of points received, the indices for multiple-choice, short-answer, and open-response questions are placed on a similar scale; the index ranges from 0 to 1 regardless of the question type. Although this index is traditionally described as a measure of difficulty (as it is described here), it is properly interpreted as an "easiness index" because larger values indicate easier questions. An index of 0 indicates that no student received credit for the question, and an index of 1 indicates that every student received full credit for the question.

Item-test Correlations

Within classical test theory, these relationships are assessed using correlation coefficients that are typically described as either item-test correlations or, more commonly, discrimination indices. The discrimination index used to analyze MCAS multiple-choice items and short-answer items, which are scored 0 or 1, was the point-biserial correlation between item score and a criterion total score on the test. For open-response items, item discrimination indices were based on the Pearson product-moment correlation. The theoretical range of these statistics is from −1 to 1, with a typical range from .3 to .6. Discrimination indices can be thought of as measures of how closely a question assesses the same knowledge and skills assessed by other questions contributing to the criterion total score. That is, the discrimination index can be interpreted as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For MCAS, appropriate criterion scores were selected based on item type and function (common or matrix). The selected criterion scores are provided in Table 13-1. For example, the criterion score for common open-response and short-answer items was the total score on all common multiple-choice, open-response, and short-answer items. (MDOE, 1999, p. 78).

The very next page of the *1998 Technical Report* presents a summary table of the average difficulty and discrimination of different question types for each subject and grade tested on the 1998 MCAS. This table (Table 13-2 from MDOE, 1999, p. 79) is reproduced in Table 2 below.

369

What these results reveal is that almost all items selected for inclusion on the MCAS tests were ones which showed item difficulties in the range of 0.35 to 0.65, meaning that between 35% and 65% of test-takers answered these items correctly. Similarly, almost all items selected for inclusion on the MCAS tests showed item discriminations in the range of 0.30 to 0.60.

Table 2
Average Difficulty and Discrimination of Different Question Types
For Each Subject and Grade, MCAS

	Reading			Mathematics			Science & Technology		
	n	Diff	Disc	n	Diff	Disc	n	Diff	Disc
Grade 4									
All MC	124	0.61	0.36	81	0.61	0.34	98	0.64	0.32
Common MC	28	0.61	0.38	21	0.61	0.35	26	0.65	0.32
Matrix MC	96	0.61	0.36	60	0.62	0.33	72	0.64	0.32
Short Answer	-	-	-	17	0.5	0.37	-	-	-
Open response	29	0.44	0.49	18	0.47	0.55	18	0.46	0.43
Grade 8									
All MC	124	0.66	0.37	81	0.54	0.35	98	0.6	0.32
Common MC	28	0.68	0.34	21	0.58	0.36	26	0.57	0.29
Matrix MC	96	0.66	0.37	60	0.53	0.35	72	0.62	0.33
Short Answer	-	-	-	17	0.52	0.49	-	-	-
Open	29	0.47	0.56	18	0.38	0.64	19	0.37	0.54
Grade 10									
All MC	128	0.64	0.35	111	0.45	0.32	128	0.56	0.3
Common MC	32	0.66	0.34	27	0.55	0.37	32	0.58	0.29
Matrix MC	96	0.63	0.35	84	0.42	0.3	96	0.55	0.3
Short Answer	-	-	-	17	0.41	0.46	-	-	-
Open response	32	0.43	0.59	32	0.24	0.62	32	0.22	0.52

These results make it clear the MCAS has been developed using norm-referenced test construction procedures. Items were selected for inclusion on the MCAS in terms of item difficulty and discrimination. This means that items which all students tended to answer correctly when the items were pilot tested are excluded from operational versions of the MCAS tests. Why? Because items that all students answer correctly (or all answer incorrectly) have no power to discriminate among test takers. Standard textbooks on testing point out that when 90% of test takers answer an item correctly, the maximum index of discrimination it may have is 0.20 (Anastasi, 1982, p. 208). As a result, items that show passing rates of 90% or greater (or discrimination indices less than 0.20) were systematically excluded from the common pool of operational MCAS from which students' scores are derived.

Appendix B to the *1998 Technical Report* presents details of item statistics for almost all MCAS items administered in 1998 (for reasons not explained item statistics for the 1998 grade 4 math test are not included.). The Appendix B data tables show that there were at least nine MCAS matrix items that showed difficulties of 0.90 or more — meaning that at least 90% of students taking the items answered the items correctly). The 1999 MCAS Technical Report contains no analogous appendix showing details for item statistics for the operational MCAS tests for 1999. Hence we cannot be sure whether all of the easy items pilot tested in 1998 (that is, ones which were answered correctly by 90% or more of students who took them in 1998 on a pilot basis) were excluded from the operational MCAS tests for 1999. Nonetheless one direct comparison of results from the 1998 and 1999 technical reports clearly shows that pilot items answered correctly by large proportions of students in 1998 tended to be excluded from the 1999 operational tests.

Table 3 shows the average discrimination of 1998 matrix multiple choice items and 1999 common multiple choice items for the MCAS tests administered at grades 4, 8 and 10 in those years. As can be seen, the average discrimination for the 1999 common (operational) tests are consistently higher than the average discrimination for the 1998 matrix items, which consisted mainly of items being pilot tested for future operational use. This contrast shows that developers of the 1999 MCAS tests clearly tended to select items with higher rather than lower discrimination. And this means that they would have systematically discarded items answered correctly during the 1998 pilot test of matrix items (recall that when 90% of test takers answer an item correctly, the maximum index of discrimination it may have is 0.20).

Table 3

Average Discrimination for MCAS Multiple Choice Items,
Matrix Items 1998 and Common Items 1999

	Matrix Items 1998			Common Items 1999		
	Reading	Math	Science & Tech.	Reading	Math	Science & Tech.
Grade 4	0.36	0.33	0.32	0.37	0.39	0.33
Grade 8	0.37	0.35	0.33	0.39	0.41	0.38
Grade 10	0.35	0.30	0.30	0.39	0.39	0.38
Sources: Table 13-2 MDOE, 1999; Table 15-2, MDOE, 2000.						

Selecting items in terms of item difficulty and discrimination is standard practice for norm-referenced tests of aptitude, ability and achievement. The Scholastic Aptitude Test (now called the SAT I), for example, was constructed with item specifications calling for most items to have biserial correlations of item discrimination in the range of 0.30 to 0.70 (Donlon 1984, p. 48). Here is how one authority on standardized testing, Anne Anastasi, described the rationale for selecting items in terms of difficulty:

> Most standardized ability tests are designed to assess as accurately as possible each individual's level of attainment in the particular ability. For this purpose, if no one passes an item, it is excess baggage in the test. The same is true of items that everyone passes. Neither of these types of items provides any information about individual differences. Since such items do not affect the variability of test scores, they contribute nothing to the reliability or validity of the test. The closer the difficulty of an item approaches 1.00 or 0, the less differential information about examinees it contributes. Conversely, the closer the difficulty approaches 0.50, the more diffferentiations the item can make. (Anastasi, 1982, p. 193)

Anastasi goes on to point out that for different kinds of testing, that is other than norm-referenced standardized testing, different kinds of test selection strategies would be appropriate:

> [M]astery testing is often associated with criterion referenced testing. If the purpose of the test is to ascertain whether an individual has adequately mastered the basic essentials of a skill or whether he or she has acquired the prerequisite knowledge to advance to the next step in a learning program, then the items should probably be at the 0.80 or 0.90 difficulty level. Under these conditions we would expect the majority of those taking the examination to complete nearly all items correctly. Thus the very easy items (even those passed by 100% of the cases), which are discarded as nondiscriminative in the usual standardized test, are the very items that would be included in a mastery test. (Anastasi, 1982, p. 193)

Lake Woebeguaranteed

High stakes testing is politically popular in Massachusetts as in other states. By high stakes testing, I refer to the use of standardized test results in isolation to make decisions about students or schools. Such use is contrary to professional standards regarding test use. (See, for example, the statement of the American Educational Research Association available at http://www.aera.net/about/policy/stakes.htm)

Regarding use of test results to make decisions about individuals, decades of research regarding college admissions testing show that it is far more sound (more valid and with smaller adverse impact on minorities and females) to make decisions flexibly using test scores, grades and other information rather than to make decisions mechanically based on test scores alone (Linn, 1982; Willingham, Lewis, Morgan & Ramist, 1990, Haney, 1993).

In the analyses presented above I have shown the folly of using annual MCAS test results in isolation to rate schools. Using data from MCAS testing in Massachusetts, we have seen the "volatility" of annual school average MCAS scores. Next I discussed three broad reasons why MCAS score averages are such poor indicators of school quality; namely, because groups of students tested can vary from one year to the next; because the MCAS tests are of dubious technical quality; and because the MCAS tests have been constructed using norm-referenced test construction techniques whereby items are selected for inclusion in terms of item difficulty and discrimination.

It is the latter practice that brings us to Lake Woebeguaranteed. When test items are selected in terms of how well they discriminate among individual test takers, this means that the test results will tend to have little power to differentiate among schools (Madaus, Airasian & Kellaghan, 1980). And when items are systematically excluded from operational versions of the tests when more than 70% of pilot test students can answer them correctly, this flat out guarantees continuing failure on the tests. It also may help to explain Bolon's finding as to why grade 10 MCAS math school average scores are so strongly correlated with community income. The MCAS math and English language arts tests have been constructed using the sort of techniques used in building tests of verbal and quantitative aptitude. It has long been established that such aptitude test results are consistently associated with family socio-

economic status (see Donlon 1984, p. 183 for just one piece of evidence on this point).

In closing let me mention two additional reasons why the use of results from tests like the MCAS to make high stakes decisions about schools and students is fundamentally ill-conceived. Recent research by Russell and colleagues (Russell & Haney, 2000; Russell & Plati, 2001) has shown that "low-tech" tests (that is, paper-and-pencil tests in which students have to write longhand), in general, and the MCAS in particular, seriously underestimate the skills of students accustomed to working on computers.

Finally if nothing else, the recent expose in the *New York Times* of widespread errors in test scoring and reporting in the testing industry (Henriques & Steinberg, 2001; Steinberg & Henriques, 2001) should make clear how unwise it is to make important decisions based on test scores in isolation. In just the last few years, virtually every major test developer has been found to have committed a major blunder, as a result of which, for example, students were wrongly forced to attend summer school, students were mistakenly denied high school diplomas and schools were incorrectly sanctioned for performance.

Notes

[1] I would like to thank Anne Wheelock, Damian Bebell, Ron Nuttall, Craig Bolon and five members of the EPAA editorial board for comments on a previous version of this article. Nonetheless, as always should be the case, responsibility for the content and conclusions of the article is solely that of the author.

[2] Several reviewers of a previous version of this article suggested that I ought to make clear that the issues discussed herein are related to longstanding statistical issues relating to the measurement of change, regression to the mean, and sampling theory. For such suggestions, I am grateful, even though I decided not to go into these more general issues in this article.

References

Ananstasi, A. (1982). *Psychological testing* (5th Edn.) New York: Macmillan.

Bloom, B. (1964). *Stability and change in human characteristics*. New York: John Wiley & Sons.

Bolon, C. (2001, October 16). Significance of Test-based Ratings for Metropolitan Boston Schools. *Education Policy Analysis Archives, 9*(42). Retrieved October 18, 2001 from http://epaa.asu.edu/epaa/v9n42/.

Bolon, C. (2002, January 28). Response to Michelson and to Willson and Kellow. *Education Policy Analysis Archives, 10*(10). Retrieved February 4, 2002 from http://epaa.asu.edu/epaa/v10n10/.

Cannell, J. J. (1987). *Nationally normed achievement testing in America's public schools: How al 50 states are above the national average*. Daniels, WV: Friends for Education.

Council of Chief State School Officers, (December 1998). Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance. Washington, DC: CCSSO, State Education Assessment Center.

Donlon, T. (Ed.) 1984. The College Board Technical handbook for the Scholastic Aptitude Test and achievement tests. New York: College Entrance Examination Board.

Gallagher, E. (2001). An analysis of problems on two 10th grade MCAS math tests http://www.es.umb.edu/edg/MCAS/mcasproblems.pdf, accessed December 3, 2001.

Haney, W, Madaus, G. & Lyons, R. (1993). *The Fractured Marketplace for standardized testing*. Boston: Kluwer.

Henriques, D. & Steinberg, J. (May 20, 2001). Right answer, wrong score: Test flaws take toll, *New York Times*, p. 1. http://www.nytimes.com/2001/05/20/business/20EXAM.html.

Kane, Thomas, and Staiger, Douglas (April 2001). Volatility in school test scores: Implications for test-based accountability systems. Working paper of the National Bureau of Economic Research. (http://www.nber.org/papers/w8156).

Kurtz, M. (September, 2001) The MCAS: High stakes testing in Massachusetts: An overview of the debate. Draft. The Rappaport Institute for Greater Boston, The John F. Kennedy School of Government, Harvard University

372

Lindsay, J. (2001). Students find mistake in math portion of MCAS test. The Associated Press, 5/22/01 6:51 PM. (http://www.masslive.com/newsflash/index.ssf? /cgi-free/getstory_ssf.cgi? g0102_BC_MA--MCASMistake&&news&newsflash-massachusetts, accessed 6/10/2001).

Madaus, G., Airasian, P. and Kellaghan, T. (1980). School effectiveness: a reassessment of the evidence. New York: McGraw-Hill.

Massachusetts Department of Education (October, 1999). Massachusetts Comprehensive Assessment System 1998 Technical report. (MCAS 98tech_report_full.pdf, downloaded, February 24, 2000)

Massachusetts Department of Education (November, 2000). Massachusetts Comprehensive Assessment System 1999 Technical Report. (MCAS 99tech_report_full.pdf, downloaded, September 13, 2001)

Michelson, S. (2002, January 28). Reactions to Bolon's " Significance of Test-based Ratings for Metropolitan Boston Schools". Education Policy Analysis Archives, 10(8). Retrieved February 4, 2002 from http://epaa.asu.edu/epaa/v10n8/.

Russell, M. & Haney (2000). Bridging the gap between technology and testing. Education Policy Analysis Archives, 8 (41). Retrieved May 3, 2002 from http://epaa.asu.edu/epaa/v8n41/

Russell, M. & Plati, T. (2001). Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment. Teachers College Record On-line. Available at http://www.tcrecord.org/Content.asp?ContentID=10709

Steinberg, J. & Henriques, D. (May 21, 2001). When a test fails the schools, careers and reputations suffer, New York Times, p. 1. http://www.nytimes.com/2001/05/21/business/21EXAM.html.

Willson, V. L. & Kellow, T. (2002, January 28). Confusing the messenger with the message: A reponse to Bolon. Education Policy Analysis Archives, 10(9). Retrieved February 4, 2002 from http://epaa.asu.edu/epaa/v10n9/.

About the Author

Walt Haney
Center for the Study of Testing, Evaluation and Educational Policy
Campion Hall 323
Lynch School of Education
Boston College
Chestnut Hill, MA 02467

617-552-4199
617-552-8419 (Fax)
Email: haney@bc.edu
Home page http://www2.bc.edu/~haney/

Walt Haney, Ed.D., Professor of Education at Boston College and Senior Research Associate in the Center for the Study of Testing Evaluation and Educational Policy (CSTEEP), specializes in educational evaluation and assessment and educational technology. He has published widely on testing and assessment issues in scholarly journals such as the *Harvard Educational Review, Review of Educational Research*, and *Review of Research in Education* and in wide-audience periodicals such as *Educational Leadership, Phi Delta Kappan*, the *Chronicle of Higher Education* and the "Washington Post." He has served on the editorial boards of *Educational Measurement: Issues and Practice* and the *American Educational Research Journal*, the *Journal of Technology Learning and Assessment* and on the National Advisory Committee of the ERIC Clearinghouse on Assessment and Evaluation. Among other publications he has authored or co-authored a number of previous publications in *Education Policy Analysis Archives* (http://epaa.asu.edu/epaa/v5n3.html, http://epaa.asu.edu/epaa/v7n4/, http://epaa.asu.edu/epaa/v8n19.html, and http://epaa.asu.edu/epaa/v8n41/)

Appendix on Data Sources and Data File

> The reader may download the datafile in the form of
> an Excel Spreadsheet. This spreadsheet is
> approximately 1000K (one megabyte) in size. It
> should load into most versions of Excel.

The MCAS data used in this study were drawn from three CD data disks issued by the Massachusetts Department of Education entitled "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of May 1998", "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of May 1999" and "School, District and State MCAS Results, Grades 4, 8 and 10, Tests of Spring 2000" and an Excel file named "MCAS 2001 ctable.xls" downloaded from http://boston.com/mcas/ on November 9, 2001. For some odd reason, this file was not available on the Massachusetts Department of Education web site when the 2001 MCAS results were publicly released on November 8.

The primary variables used in merging the four sets of grade 4 MCAS math results were district and school code numbers. However there was a small number of cases in which ID codes (or school names) were not the same, but data from the four years were still merged to represent a single school "case." Reasons for this were as follows. In Beverly, the Mckay school has a school ID number of 35 in the 1998 data set, but an ID number of 37 in the 1999, 2000 and 2001 data sets. Similarly in the 1998 data set two schools in Easton (the Fl Olmstead and HH Richardson schools have a different school ID code number in the 1998 data set than in data sets for subsequent years. In general, data were merged based on school ID code numbers and names, though across the four data sets, there were numerous variants in school names. In several instances cases which had both different names and school ID across the four data sets were treated as a single school. These were four instances in which there was only a single school in a town, but which had roughly the same number of students tested across the across the four years of MCAS results. In the 1998 data set, the town of Chesterfield was reported to have one school, named Center School, in which 12 students were tested. In the 1999, 200 and 2001 data sets, the town of Chesterfield was also reported to have one school, but one named "NEW HINGHAM REGIONAL ELEM," with 14, 19, and 19 students reported as tested in the latter three years. In cases such as this, that is a town with only a single school with roughly the same number of students tested across the four years of MCAS results, data across the four years were considered to represent a single school case.

The one school in Holliston with grade 4 students tested was treated as a single case even though, the name of the school in 1998 and 1999 was "Flagg Adams Middle" but "MILLER SCHOOL" in 200 and 2001. In Dover, the name of the one school with grade 4 students tested was "CARYL SCHOOL" in 1998, 1999 and 2000, but CHICKERING" in 2001. The one school in Medford was named "GREEN MEADOW SCHOOL" in 1998–2000, but" FOWLER MIDDLE" school in 2001.

Also, two schools in Malden were treated as single cases because their records across the four years of data contained identical school names, even though the school code numbers varied by one digit. Similarly, the" KINGSTON ELEMENTARY" school was treated as a single case even though the ID code listed for it changed from 5 in 1998, 1999 and 2000 to 20 in 2001.

Because of such issues, in the data set accompanying this article, I have included the original district and school codes for all four years of MCAS results, plus the original variable labels for all four years of MCAS results. This will allow anyone undertaking secondary analysis to make their own decisions about the cases described above.

Education Policy Analysis Archives

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Policy and Practice:
Restructuring Teachers' Work

Lisa Kirtman
California State University, Fullerton

Abstract

Despite repeated attempts to reform schools, teachers' work has remained surprisingly stable. The
purpose of this study was to investigate implementation of a state-funded restructuring initiative
that intended broad changes in teachers' professional roles. Sponsors of the founding legislation
reasoned that changes in teachers' roles would contribute to higher student achievement. This
study examined the question of whether and how this program of comprehensive whole-school
change promoted changes in teachers' roles in school governance, collegial relations, and the
classroom. Further, the study traced the relationship of these changes to one another, and
weighed the likelihood that they had the capacity to affect core educational practices.
Theoretically, this study is situated in the available literature on teachers' collegial relations;
participation in shared decision making; and classroom roles, relationships and practice. Three
elementary schools served as the sites for intensive qualitative data collection completed over a
two-year period. The schools differed in geographic location (two urban, one rural), but all
enrolled a racially, ethnically and linguistically diverse population of students, and more than
half of the students in each school qualified for free or reduced price lunch. The study resulted in
multiple types and sources of data on teachers' professional roles, including: observations in
classrooms, collegial interactions, and governance situations; interviews with teachers (including
teacher leaders), parents, administrators, and students; and documents pertaining to the
restructuring plans and process. Findings show that changes in the three areas were achieved
unevenly in the three schools. All three schools introduced changes in classroom practice and
roles, ranging from the adoption of multi-age classrooms to more modest innovations in
curriculum or instruction. In only one case were changes in professional roles outside the
classroom organized to support and sustain classroom changes. Two of the three schools
introduced changes in staff organization (teacher teams) and leadership (governance
committees), but under-estimated the professional development and other supports that would in
turn support changes in classroom practice. Altogether, it appears unlikely that the observed
changes in professional roles were sufficiently well established and connected to affect core
educational practice in the long run.

Introduction

Despite repeated attempts to reform schools, teachers' work has remained surprisingly stable. From 1880 to the present, little has changed in the organizational structures, instructional practices, and authority structures of teachers' work (Cuban, 1993). Some authors (Weick, 1976; McNeil, 1988) theorized that this stability is due to the fact that school governance has been situated in the hands of individuals external to the classroom. Lortie (1975) has argued that it is due to the fact that much of teachers' work inside the classroom has been largely independent and individually-controlled. Still others (Rosenholtz, 1991; Cuban, 1993) have argued that teacher-centered instruction is the culprit. Whatever the reason, teachers' work today remains fairly similar to that of 100 years ago; it is characterized as individual work, with the governance power situated in the hands of individuals external to the classroom, and instruction that is largely teacher-centered (Lortie, 1975; Sarason, 1982; Cuban 1993).

More recently, Elmore (1996) and others theorized that this stability in teachers' work may be due to the fact that many past reform efforts have not successfully affected the "core" of educational practice. He defines the core of education practice as the teachers and students' role in learning and school practices, and how these ideas about knowledge and learning are manifested in teaching and the classroom. The core also includes structural arrangements of the school or classrooms such as physical layout, student grouping, as well as communication among parents, teachers and staff. In short, reforms have not affected what teachers and students do when they are together.

To illustrate this point, reform reports of the late 1980s have devote relatively little attention to the implications that reform initiatives have for teachers' work, professional roles, and collegial relationships. For example, the recommendations contained in California's elementary school reform blueprint, *It's Elementary* (1992), touch on the learning environment of the classroom, diversity, and technology as well as organizational issues such as scheduling class work in larger blocks. But the report failed to consider the teacher's role in governance and only touched on some aspects of the teachers' role in classroom or collegial relationships. Thus, the prospects for changing the core of education were reduced.

In 1990, California's School Restructuring Demonstration Program envisioned comprehensive changes in teachers' professional roles that would result in more "powerful learning" for students (California Center for School Restructuring, 1993). One of the purposes of California Senate Bill 1274 (SB 1274) was to test the feasibility of large-scale systemic school reform, the hope being that the bill would affect school sites beyond the schools participating in the bill. In the final form, the bill included the following,

"The demonstration of restructuring is intended to be a five-year effort aimed at improving student learning. The demonstration centers on the goal of engaging all students in powerful learning experiences, and a rich-thinking curriculum which empowers them to become life-long learners. All students, regardless of race, ethnic, linguistic or socioeconomic background need to learn to think critically, solve problems individually or as part of a team, analyze and interpret new information, develop convincing arguments, and apply their knowledge to new situations. The demonstration invites educators to consider radical changes in the way schools and districts operate in order to create a better environment for engaging all students in powerful learning experiences and in a rich, meaning-centered curriculum" (CSBED, 1990, p. 1)

Schools were asked to create new structures and practices that included increased professional collaboration and capacity-building, a greater number of diverse stake holders in decision-making processes, improved curriculum assessment and diverse instructional strategies, increased inquiry by examining students work, and better shaped specific strategies that impact the whole school.

The bill, although generally vague in meaning, would require changes in teachers' work to be successful. This approach of changing teachers' work to change the classroom is uncharacteristic of past reform bills that have virtually ignore teachers' work. Moreover, the bill seems to touch on the issues outlined by Elmore's (1996) core theory. The call for professional collaboration and capacity building could affect the teachers and adult communication and relationships on the school level (collegial relations). The call for the inclusion of greater number of stakeholders in decision-making processes could require a change in teacher, parent and staff roles in school governance (governance). Finally, changes in curricular instructional strategies and the examination of student work could impact teacher and student roles in the classroom (classroom roles and relationships). In short, the bill would require changes in teachers' work in the areas of collegial relations, governance and classroom roles and relationships.

Instead of once again studying why teachers' work has remained stable or why the core has not changed, SB 1274 allowed me to look at a bill that would require changing teachers' work in order to change the core. Using schools that restructured according to Senate Bill 1274, I investigated the following questions: 1) Under what, if any, conditions can restructuring promote changes in teachers' professional roles and practices? 2) Do these changes have the capacity to affect the "core" of educational practice?

The Schools[4]

This section offers a description of each of the focal schools, showing how they were positioned to undertake

comprehensive restructuring, and recording the choices they made in the three areas of collegial relations, governance, and classroom roles, relationships and practice. Schools made restructuring choices based in part on SB 1274's theory that altering professional roles and relationships would improve academic achievement. To illustrate this point, I will highlight the following issues: Restructuring effort of each school; Teacher, administrative, parent and student relationships; Externalities that may have affected restructuring.

Web Magnet School

Web is a science and technology magnet school in a small urban district. Web is the smallest of the three case study schools. Web Magnet school was created in 1990. As a magnet school, the school is designed to serve the district's high achieving students. The school is located across the highway from the poorer neighborhood from which it draws most of its students, but the school itself is located in a middle to upper income area. Web does not screen its entering classes of kindergartners, but students transferring to Web from other schools are tested for high achievement levels in math and language arts. Web's students routinely perform better on standardized tests than the rest of the district schools. The wait list for Web enrollment is very long; one parent reported waiting almost two years before her daughter could enter the school. Web serves 280 Pre K through 8th grade students. While the district is approximately 65% Latino and 35% African American, Web's student body is 62% African American, 30% Latino, 3% White, and 3% Asian, Pacific Islander, and Filipino (School Documents, Fall 1995). Twenty-three percent of the students are classified as LEP, which has grown from 7% in three years (School Documents, Fall 1995). Fifty one percent of the students qualify for free or reduced meals (School Documents, Fall 1995). A teacher gave us the following description of the school:

> T: It opened as a magnet. It opened in Fall of '88 as a science and technology magnet. It opened under the superintendent's guidance, [her] dream/vision and parent wish that...it was, you see [a new school] had just started not too long before that and so all of the more mainstream families in the community were going out to the surrounding districts so, [the superintendent] wanted to have something that would keep these people in the community because it's not, and I...any family that asks me and there are many that trust me now, enough to respect the response, when they're talking about sending their child out, I will tell them, don't do it, don't do it because it's not a kind place for children in these other schools (Black, Teacher, interview, Spring 1997).

Web's staff consists of 10 classroom teachers, a science teacher, a principal, a computer teacher and a librarian. Other part time adults on site include an *I Have a Dream* coordinator, a *Reading Recovery* teacher, a day care teacher, and tutors from a neighboring university. The staff is 71% white and 29% African American (School Documents, Fall 1995). This school has had two principals since the restructuring process began and both are African-American women. The science teacher is the only male staff member. Web's staff has experienced high turnover, and consequently, there are no staff members remaining who were at the school when the original 1274 grant was written.

Teachers' responsibilities, in addition to teaching, include yard duty, bus duty, computer room duty and fellowship committee. Yard, bus, and computer duty consists of the supervision of children in the assigned areas before and after school. The fellowship committee is in charge of "school parties that keep the morale of everybody up" (Dole, teacher, interview, Fall 1995).

As part of their restructuring effort, Web teachers keep their classes for two years in a row (cycling), moving back and forth each year between two grades. For example, a 5th grade teacher will move with her class to 6th grade. Once the 6th graders move to the 7th grade, the teacher goes back to 5th grade and process begins again. In addition to cycling, Web focused its classroom restructuring efforts on the use of Integrated Thematic Instruction (ITI) as well as Brain Compatible Instruction in all of the classrooms (see Figure 1).

BEST COPY AVAILABLE

ITI and Brain Compatible Theories
(Taken from Susan Kovalik and Associates)

The ITI model begins with an understanding of six basic concepts coming from bodybrain research:

1. Emotions are the gatekeeper to learning and performance.
2. Intelligence is a function of experience.
3. Humans in all cultures use multiple intelligences to solve problems and to create products.
4. The brain's search for meaning is a search for meaningful patterns.
5. Learning is the acquisition of useful mental programs.
6. Personality - one's basic temperament - affects how a learner takes in information, organizes and uses it, and orients him/herself with respect to the world and other learners.

Once fully understood, Kovalik's ITI model leads educators to the eight brain-compatible elements as a guide for applying the research through thoughtfully written curriculum and carefully selected teaching strategies:

1. Absence of threat
2. Meaningful content
3. Choices
4. Adequate time
5. Enriched environment
6. Collaboration
7. Immediate feedback
8. Mastery (application level)

In an ITI classroom, students know what they are studying and why. The focus is on developing student understanding of important concepts, such as change, through curriculum that begins with a location or event in the student's world. As students investigate and conduct research to answer the big question, "What's going on around here?" the teacher ensures that state and local learner goals are addressed. At all times, the ITI teacher has answers for the pivotal questions, "So what?" and "Why do we have to learn this?" The teacher can answer Susan Kovalik's guiding questions, "What do you want them to understand?" and "What do you want students to do with it?"

Figure 1. Web Restructuring Philosophy

One thing that sets Web school apart from the two other schools is the union involvement. Several members of the staff were union leaders, and many teachers were very involved in union activities.

Olive Grove Elementary School

Olive Grove Elementary School serves 576 kindergarten through 6th graders in a largely rural area in Northern California. Between 1990 and 1992, the school's Hmong and Mien student population rose from almost nothing to 29 % of the student body. The school is still shaping its Limited English Proficiency (LEP) program in the wake of this population change. Fifty-five percent of the students are white; 7% are Latino; and 7% are African American (School Documents, Fall 1995). Fifty-one percent come from families with income low enough to qualify for Aid to Families with Dependent Children (AFDC) (School Documents, Fall 1995). The school is in the process of applying for Chapter One status based on the high number of students who qualify for free and reduced lunch. When asked about the history of the school a teacher said,

> T: I think it's important for you to know for your study, do you have any history of where we are. We were pretty well traditional, felt a real, real need to change. Our population was changing rapidly, not only in terms of a huge influx of immigrants, largely Asian immigrants with no English background, we had that, this neighborhood area that we take in, is a very low economic area. Our welfare rate then was high, it's much higher now. We have a huge incidence, at any given time, I have one, two or three kids who have one or both parents in jail. Right now, there are two, mostly drug related problems (Darvy, teacher, interview, Fall 1995).

Olive Grove's teaching staff consists of 23 classroom teachers, 14 aides, and a resource teacher (School Documents, Fall 1995). Three of the 23 classroom teachers are male (School Documents, Fall 1995). There are Hmong bilingual aides. The majority of the staff is white, and the principal is a white male.

As part of their restructuring initiative, the school changed its classrooms to form multiage/multiyear groupings. Classrooms are either self-contained kindergarten, 1/2 combinations, or 3-6 clusters, which consist of approximately 7 third graders, 7 fourth graders, 7 fifth graders, and 7 sixth graders. Children stay with the same teacher and classmates in each of the groupings, culminating in four years with their 3-6 grade teacher. In addition, teachers are grouped in three K-6 grade teams, with approximately seven classrooms making up each team. The school has also adopted class meetings as an instructional change. Teachers are required to hold class meetings everyday. Class meetings begin with students and teacher sitting in a circle. They give complements, they discuss problems, and they discuss class business.

The school has experienced little staff turnover since the grant began; 15 of the 23 teachers working when the grant was awarded in 1991 are still on site, as is the principal. The average years of teaching experience for the staff is 10 years.

Trent Charter School

Trent is located in a low-income section of a large metropolitan area in southern California. The school serves 1146 students in Pre K through 6th grade. Ninety-six percent of the students are Latino; 3% are African American; and the rest are Asian, Filipino, American Indian, and White (School Documents, Fall 1995). Eighty-one percent of the students are classified as Limited English Proficient (LEP), and the school conducts many of its classes and almost all of its yard and lunch activities in Spanish (School Documents, Fall 1995). Ninety-six percent of the students have family incomes low enough to qualify for free meals (including both breakfast and lunch served at school) (School Documents, Fall 1995).

The 126 staff members include 45 classroom teachers, 36 paraprofessionals, and 5 administrators/coordinators (School Documents, Fall 1995). Seventy-two percent of the certificated staff has level "A" fluency or a bilingual credentials (School Documents, Fall 1995). The principal is a Chinese American woman who is fluent in Spanish.

Trent began its reform efforts when the current principal arrived and changed the school governance structure to site based decision-making. Several major grants followed her arrival, including a United Way grant for a parent center, a Healthy Start grant, an RJR Nabisco Next Century grant, and the SB 1274 School Restructuring grant. The school became a California Charter School in 1993, and staff points to this change as the most significant for the school. SB 1274 is, therefore, one piece in a much larger school change effort at Trent.

Trent divides its students into 45 single grade classrooms, which are designated Limited English, English Only, Bilingual (a mix of the first two), Transitional, and GATE. Parents may request the type of classroom they would like to enroll their child in, provided there is room. Parents often request English classrooms, despite staff attempts to convince them of the worth of primary language instruction. The school attempts to transition all its students to English classrooms by the end of the third grade. The school operates on a year round calendar, and last year they used their restructuring money to fund twenty extra pupil days, increasing attendance days from 180 to 200 per year (ESY Days).

School governance is carried out by eight governance committees. Teachers are required to serve on one committee, and must rotate every two years. Each committee must also have a parent member.

The school has focused its classroom restructuring efforts into making sure the writing process is taught in all 1-6th grade classrooms. Also as part of the restructuring effort, the Parent Center was created. The school philosophy is that

the school should be the center of the community. The Parent Center is open beyond school hours; and provides clothes, food and English language training; and a referral service for other needs. In addition, in 1997, the school began work on what they call the "Village" which will include a library for public use, a supply store and a teacher-training center on campus. The Center and Village will be run entirely by parents. A parent described the affects of the Parent Center as,

> They've (parents) benefited because the Center helps with family problems. Sometimes we see kids in the yard that don't get along with the others, that fight a lot, so we refer them to the counselor at the Center. She talks with the child and contacts the family. Quite often, families come here before a big problem arises. Sometimes they need medical help, and the Center can refer them to various places. Sometimes they need financial help or counseling. So in this manner the Center is helping the kids and the school (Donner, parent, interview, Fall 1995).

Assumptions about Change

Based on the previous description and past research, many people would make assumptions regarding the possibility of successful change at each of these schools.

- Assumptions about Web: Being such a small school, one might assume that establishing relationships and creating communication channels, training and ensuring that reforms are in place would be easily achieved.
- Assumptions about Olive Grove: Based on the size of the school and the low teacher turnover, one might assume that Olive Grove's situation would be conducive to creating strong trusting collegial relations, general communication, governance change and classroom change. In addition, the multiyear configuration should provide an environment that creates strong parent/teacher relations.
- Assumptions about Trent: A large school might be assumed to have difficulty with communication, relationships, consensus and any type of wide reaching change.

But what I found is that none of the assumptions held true.

Models of Change

At the beginning of this investigation, I posed the question: Did restructuring promote changes in professional roles and practices that have the capacity to change what teachers and students do when they are together?

To determine whether a change has taken place, I will first define professional roles and practice. Professional roles and practice are split into two categories 1) inside classrooms and 2) outside classrooms (See Figure 2 below.).

Teachers' Professional Roles

Inside Classrooms
(Classrooms)

- Teacher/student relations
- Teaching techniques and strategies

Outside Classrooms
(Collegial relations and Governance)

- Committee work
- Other school responsibility
- Teacher/Teacher relations
- Teacher/Administrator relations
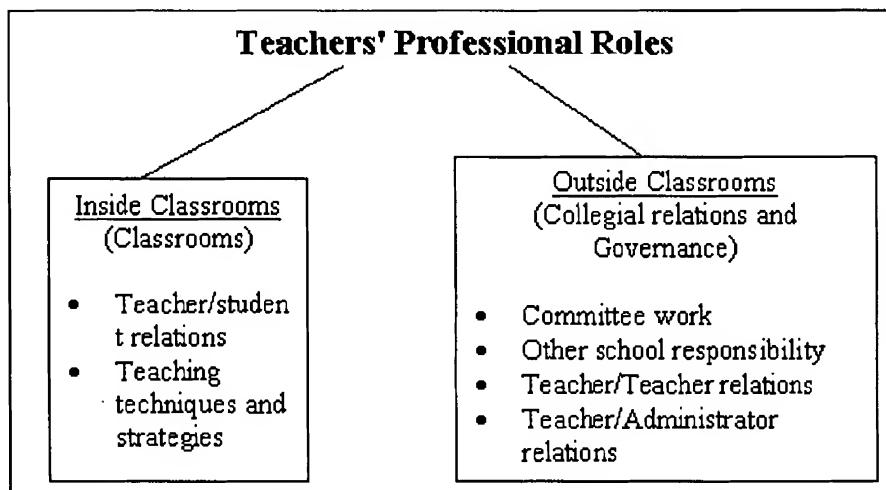
Figure 2. Professional Roles & Professional Practice

At Trent, I found that teachers' roles have been significantly altered through work on committees and in clans. Teachers run all aspects of the school including peer evaluations. But, has there been a change in what teachers and students do when they are together? The one practice that has clearly been altered at Trent is the addition of the writing process. Every teacher uses the writing process in his or her classroom at Trent. For example, in one sixth grade class, the students participated in writers' workshops (one form of the writing process) everyday that I observed.

They publish their writing on the classroom and computer lab computers (classroom observation, Fall 1995 & Spring 1996). Moreover, during an observation of the assessment committee, I observed evidence of the writing process in all 20 rooms that I visited. Teachers were either working on the process when we entered or there was evidence of its use through student and teacher work posted on the walls (classroom observation, Spring 1996).

Also a teacher said, "we push for writing process now. All they do is write, write and write in my class" (Grandville, teacher, interview, Spring 1997). To further illustrate this point, the two sixth grade focus students both said that their favorite subject was Writer's Workshop, which employs the use of the writing process.

Maybe more important than the fact that the changes occurred is the fact that the changes in professional roles seem directly related to the changes in the classroom. At Trent every teacher agreed on the writing process as a focus. Each committee sought a way to affect it. Each clan made sure every teacher was trained in it, and the assessment committee made it one of their focal points to look for when they observed classrooms. If the assessment committee did not find evidence of the writing process in classrooms, the clan was notified, and the clan made professional development and sharing of materials in that area a priority. The following diagram summarizes the change process at Trent.

Trent's Change Process

Decisions made together about reform lead to

Outside the Classroom

- Committee Work
- Clan Meetings
- Peer review
- Team Teaching

Inside the Classroom

The Writing Process

Feedback Loop: evaluation and assessment
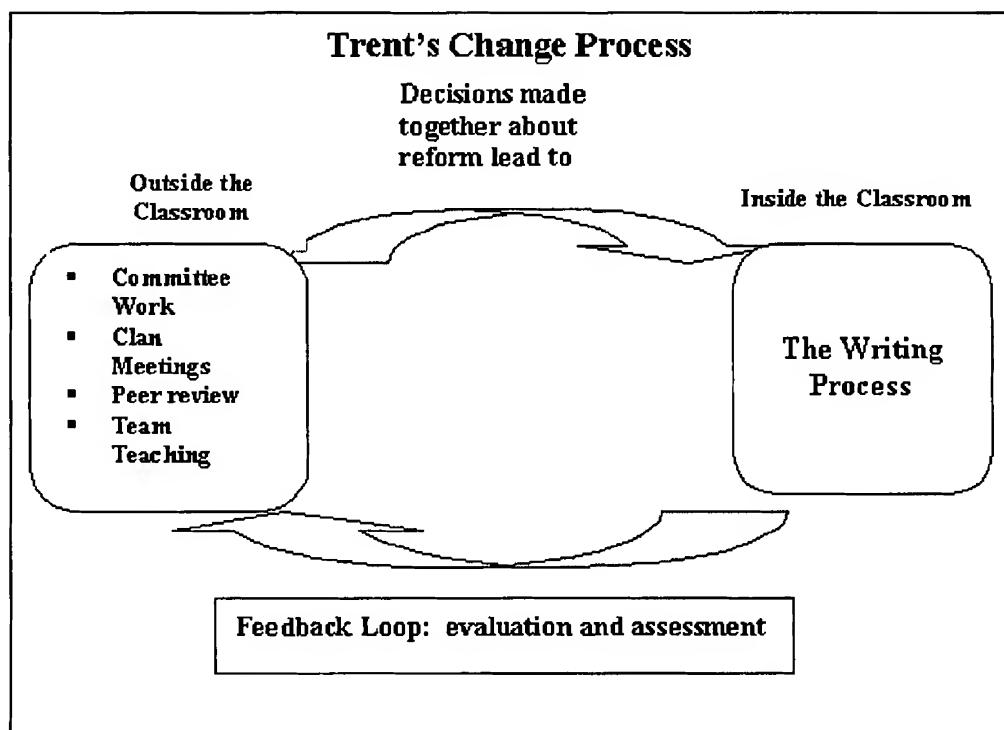
Figure 3. Trent Change Process

At Olive Grove, we also found changes in professional roles and practice, but, unlike at Trent, the changes occurred first and most clearly in the classrooms. As the first step in Olive Grove's restructuring process, they moved to a multiage/multiyear configuration. In moving to this structure, teachers' roles and practices changed inside the classroom. However, these changes were not anticipated. Teachers no longer used textbooks, and teachers stopped using directed lessons because techniques such as individualized packets made the varying ages of the students easier to teach and control. In addition, to cope with the multiage configuration, teachers needed students to help each other which allowed the "student as teacher" role to arise. Teachers could not always work with all levels at once, so teachers had to move from being the one who asked the majority of questions and gave the majority of answers to one of many teachers in the room. Finally, this change in structure led Olive Grove's teachers to change their focus. They moved from a focus on curriculum to a focus on classroom management, control of behavior and socialization.

After these changes to professional roles and practices occurred inside the classroom, Olive Grove attempted to change teachers' roles outside of the classroom through the creation of action teams. However, these teams had no real power to make changes to the school, no mechanism in place to determine the teachers' learning demands, and "no vision of what to do next" (Zucker, teacher, interview, Spring 1997). Each committee had a different goal with no real connection to or affect on the classroom. They had no mechanisms in place to help teachers cope with the new roles or assess the progress of their reforms.

Although Olive Grove made bold changes in their classrooms, their efforts fell short of the ultimate goal of increasing

student achievement. The staff lost focus of the goal and began to focus on control. Since they did not have a mechanism in place to evaluate their efforts, they were unaware of the difficulties, and thus, they were unable to refocus their work.
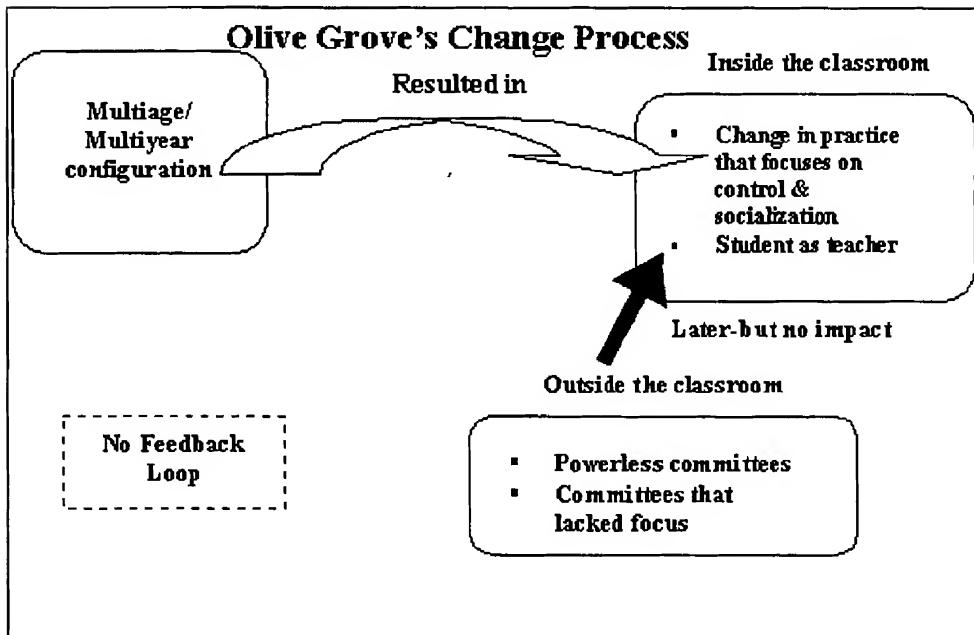

Figure 4. Olive Grove Change Process

The third school, Web, did not demonstrate clear changes to professional roles inside or outside of the classroom. Web, more than the other two schools, continues to function as a "typical" school. Teachers continued to be isolated and administrators continued to lead. However, it is important to note that Web had uniform practices across the school. The small school size enabled the administrator to spend her time checking each classroom weekly to make sure there was evidence of ITI and Brain Compatible Instruction in place. This technique is obviously one way to make change happen at a small school, but it led to low teacher morale, high turnover and superficial adoption of techniques.

Web's reforms may not have resulted in many changes for a number of reasons. It could be the lack of professional development training for new teachers, the high turnover, or the fact that as a small school, the teachers at Web already wear many hats. At a small school such as Web, teachers must take on many roles and responsibilities because they have fewer support staff than a larger school. Thus, maybe there was no shift in teachers' roles from what is typically expected because teachers' roles at Web were not truly typical to begin with.
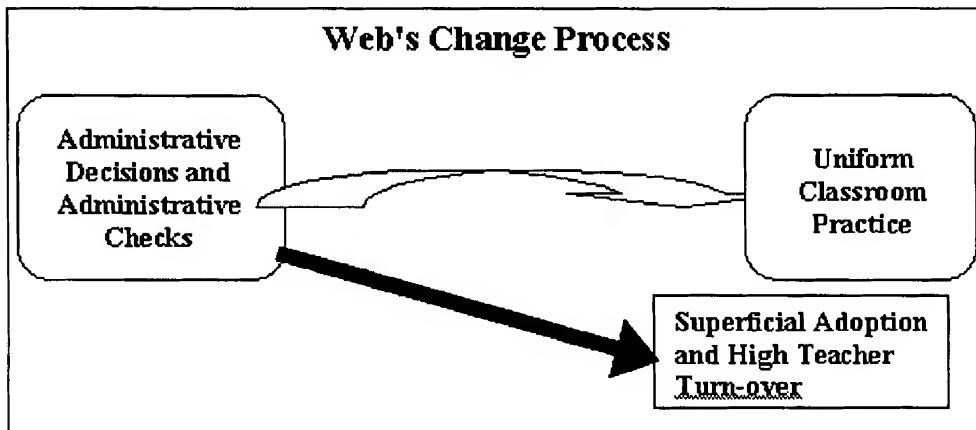
Figure 5. Web Change Process

383

There are two main points that should be taken from these models. First, changes to the classroom will be uneven, superficial or even negative if professional roles outside the classroom are not organized to support the intended reform. Support may come in the form of professional development, increased time or salary. Support should also come in the form of evaluation and assessment that are needed to make sure the reforms are taking the intended shape.

The second major point is that changes to professional roles outside the classroom are unlikely to affect classrooms unless the changes in professional roles are directly linked to classroom changes. Through examples, such as Olive Grove, I found that a school cannot just change the classroom without changing the roles and practices that control and affect the classroom. Olive Grove had no checks in place, and no way to make sure that their plan was working. They lacked the skills to cope with problems, and their communication channels were limited by friendship. The staff at Olive Grove lacked the means to assess the success and impact of the changes. They had no idea if their changes were working or if these changes were in the best interest of the children. They were unable to determine whether or not training was necessary and if it was, they had no governance structure to put the training in place. In short, there must be a mechanism in place to evaluate and assess the effects of the intended reforms.

For Web, they never began the change process as a school. They never opened lines of communication. Without communication, there was no chance for real change. They did not discuss a focus, agree on a reform or agree that a change should be made. The principal made the decisions and assessed the progress of change. This mode of operation ultimately led to superficial changes.

Although Trent seems to have all the mechanisms in place to make whole school large scale change, they did not make a fundamental change to the core. Their change was only instructional. The one thing that prohibited Trent from creating a fundamental change to the core was the identification of the problem. They identified an instructional problem. This fact leads me to my final point and that is that there has to be a clear identification of a need for a fundamental change for that type of change to occur. Even with all of the mechanisms in place for change, if the school does not believe there is a need for a fundamental change, then change will not take place.

Teachers

In research, we have a tendency to categorize people in schools. Unfortunately, these categories make us lose sight of the fact that prior experiences shape individual beliefs and understanding about events such as restructuring. What I found is that the category of teacher can be split into three different groups, and these groups experienced restructuring differently. The three groups are **teachers new to the profession, new to the school**, and **experienced**.

To further illustrate my point, this section contains profiles of three teachers to demonstrate how they experienced restructuring differently. I chose to profile three teachers from Olive Grove because this is where the data is the most complete especially in the area of new teachers.

New teachers (new to the profession and the school) were less likely to adopt their school's reform efforts. For example, at Web, all of the experienced teachers used Brain Compatible techniques in their classrooms, but the new teachers resisted this change. A teacher new to the profession said,

> I think that brain compatibility and ITI and Comer have made it more difficult for me to teach because I think I am being held to a standard that I cannot meet. Because I do not think I have been provided with any materials that I was supposed to have...I should not have to paint my room or buy a CD player or buy furniture. I feel I got a lot of political pressure at this school. To conform and change my classroom. I got a lot of negative feedback about colors and furniture arrangement (teacher, focus group, Spring 1997).

Even the experienced teacher saw the problems that new teachers (to the professional and to the school) were having:

> Um, well, uh, you know, from the perspective of the new teacher started this year. They have to do portfolios. I think that's unheard of for a new teacher to have to do a portfolio. And to write their own curriculum. I think that's unheard of. I just think so many things are just, you shouldn't hire a first year teacher and expect them to be able to do those kinds of things or to um, thematic teaching. I think it's really difficult. Or to have to have people come in and observe your classes when it's your first year. I think that's really difficult. And they do that at our school. And I don't mind because I've been teaching a while, and I don't care if people want to come in. But I think that for many people, that's really scary. It's really scary (Migdal, teacher, interview, Fall 1995).

The new teachers' (both) rooms were different from the experienced teachers at least in part due to a lack of understanding about the reform and a lack of support in the form of material and professional development.

The new teachers (both) were also more likely to have teacher-centered classrooms. At Trent,

New Teacher Profile

many teachers tried to get away from workbooks and worksheets, but a new teacher said, "Catholic school is a first year teachers dream–everything had a set of books: One book and one workbook for language and spelling and reading, hand writing and phonics. Here there are not nearly as many books" (Larson, teacher, interview, Spring 1997).

Ms. Putnam is a 1st year teacher at Olive Grove. She has 26 3-6 grade students in her classroom. The students sit in groups based on grade level. As an example, when students work on a math lesson, she takes one grade level to work with while the other groups work in their grade level appropriate textbooks. Ms. Putnam said of the multiage/multiyear configuration: "their (the other teachers) styles are a lot different than mine—I find that I need more structure—for me I think more freedom will come next year because I will only have 4-6th grade (S.97). As for committee work and taking on administrative responsibility, she said, "It's just another excuse for a meeting" (S.97)

Similarly, at Olive Grove, in new teachers' (both) classes, one was more likely to be able to tell which children were in which grades. In an experienced teacher's room, the class seemed as one (Putnam, classroom observation Spring 1997 and Oats, classroom observation, Spring 1997). At Web, the experienced teachers pointed out that to help focus on the students, "we don't teach from textbooks" (teacher, focus group, Spring 1997), but a new teacher said, "this is my first year with science and I didn't feel like there was as much material as there was in social studies. I need more textbooks" (Rosswell, teacher, interview, Spring 1997, p. 15). But again, these differences seem to be due to a lack of comfort with the reform that can be attributed to a lack of training or support.

"New to the School" Teacher Profile

This is Ms. Zucker's first year at Olive Grove, but her 4th year of teaching. Ms. Zucker has 30 3-6 grade students. These students sit in mixed groupings. Similar to Ms. Putnam, when she teaches a subject like math, the students sit with their grade levels and are taught out of a textbook. Her viewpoint on multiage/multiyear configuration is, "I definitely want to keep it," but she refused to continue rotating students through several classrooms each day. The year before her arrival, her team decided to share all students, but when she arrived, she refused to take part. She said, "It was uncomfortable for me and my students" (S 97). She said she would like to keep the action teams, although, in her opinion, they do not have much impact currently. She said she has been at a school where teachers had no say, and "once we've been where we are we wouldn't ant to go back to the old ways [of no input]" (S. 97).

Finally, new teachers were less likely to like the idea of taking on administrative work. When a new teacher at Olive Grove was asked if she felt committees were worthwhile, she responded by saying, "I feel it's just another excuse to have a meeting at 7:30 in the morning. Nothing has occurred thus far on the evaluation and planning committee" (Putnam, teacher, interview, Spring 1997). Similarly at Trent a new teacher said, "I want some say in governance, but not this much" (Larson, teacher, interview, Spring 1997).

New teachers (both) all suffered from the same difficulties no mater which school they were a part of. First, and maybe most importantly, the new teachers lacked the training in the areas necessary for successful reform. At Web, a new teacher said, "the only staff development was when we had a day where they said, "Is there anyone who needs help? Buddy up with other teachers who can help you" (Rich, teacher, interview, Spring 1997). Moreover, when a new teacher at Olive Grove was asked about staff development opportunities she said, "some are available but you know again nobody is talking or supporting, and no one is encouraging. You know I've been in places where everyone is trying to encourage you, you know if you don't get your masters then you need to get this or get that. And everybody is talking about continuing their education" (Colter, teacher, interview, Fall 1995). Nevertheless, despite this lack of training at Web, teachers were still expected to implement ITI and Brain Compatible work. Without proper training, it would be impossible to find successful implementation of the restructuring effort in the classrooms. They just did not have the know how to implement the expected changes.

Second, and this is where there is a distinction between new to the profession and new to the school teachers, was the problem with added administrative responsibilities. Some of the teachers new to the school were able to handle the administrative responsibilities and may even like them and see them as necessary. But, on the other hand, teachers new to teaching were overwhelmed by the administrative duties. First, new to the profession teachers were trying to learn how to teach and at

Experienced Teacher Profile

Ms Johanson has been teaching for 28 years at Olive Grove. She has 23 3-6 the graders in her class. Her class configuration and techniques are very much in line with the majority of the staff members at Olive Grove. Her students sit in mixed groupings. Math is taught based on the students assessed math level (therefore it is possible that a 3rd and 6th grader are completing the same work). There are few directed lessons and few textbooks in use. The students spend a lot of time teaching and helping each other. Ms. Johanson "loves" the multiage/multiyear configuration. In fact, she was one of the first to try it on the "exploration team" (F. 95). She also feels it is important to have committees. She is the chair of the assessment committee. She says, "it is hard work, but it has to be done" (S. 97) .

the same time being asking to lead. A new teacher said, "How can I lead when I don't know the school, and I am just trying to figure out my classroom" (Rosswell, teacher, interview, Spring 1997). At Olive Grove, a new teacher stated that she felt the committees were just another excuse for meetings, while a teacher who was new to the school but had three years teaching experience said,

Keep action teams? I'll definitely keep it. I think that we need to look at them again and look at how people look at them honestly and have an open honest discussion about it. Sometimes I think there's a feeling that we don't want to undo anything we've done for fear that it would be seen like fear that if left to their own devises, teachers would just go back to their old ways and we have to keep forging along. I don't think that's true about teachers. I think that once we've been where we are we wouldn't want to go back in a lot of ways (Zucker, teacher, interview, Spring 1997).

Trent did attempt to help the new teachers. Soon after they were hired, the new teachers (to the profession and school) were trained in the writing process Moreover, they have "changed and reduced some of the responsibilities of new teachers. So they don't have as much as more experienced teachers" (teacher, focus group, Spring 1997).

But even with these changes, new teachers (to the school and to the profession) in many ways were left out of the restructuring process, and thus their classrooms were left out of the changes.

Conclusion

The central lesson to be taken from this research is that teachers' roles have not only to be impacted but also supported to achieve school change. Elmore (1996) argued that to create change, you must change the teacher. My work supports his conclusion, but my work illustrates that effecting the teacher is not enough. The teacher has to be supported in very specific ways throughout the change process to create successful reform. Teachers must be supported through opportunities for professional development, through an assessment/evaluation feedback loop that allows for growth not punishment, and through incentive programs to encourage collegial relations and to reduce the stress involved in reform.

Throughout this work, I found that reforms repeatedly fell short of their intended goals due to a lack of support. At Web, one could assume that its small size would make reforms such as moving to school-based governance easy to establish, but I found that in the absence of opportunity to meet or incentive to meet even under ideal conditions, the change is likely to fail.

At Olive grove, where close friendships had been established among staff members, one would assume that collegial role changes such as communication and evaluating each other's work could be established. But again, without mechanisms in place such as school-wide evaluation and assessment that feedback to the teachers, Olive Grove could not see that their reforms were not working.

At Trent, where one would expect attempts at school-wide change to be unsuccessful due to the large size of the school, the assessment, professional development, material availability and incentives were all in place to make their intended reform successful.

Policymakers need to keep in mind there must be a balance between impacting teachers but supporting teachers to create successful reform. The following policy recommendations further highlight this important balance necessary for successful change:

- Professional Roles
- Externalities Matter
- New Teachers
- Restructuring is demanding and stressful
- Professional Development
- Teachers want a say

Professional Roles

The authors of SB 1274 asserted that changing professional roles can lead to changes in classrooms, and my work supports this claim. Trent provides us with a clear example of the importance of changing professional roles. Trent's teachers moved from only focusing on classrooms to a teacher as administrator role. This movement allowed the teachers to see the whole picture of reform. Teachers know what they need to make the reform work inside the classroom, and for the first time, they had the power outside the classroom to get the material, professional development, feedback and collegial support needed to achieve their goals. Porter (1989) argues that individual teachers know their students' better than any outside source, so teachers are in the best position to determine which techniques work best for their students. However, he also argues that many responsibilities are outside the teachers' expertise and are thus best controlled by administrators. What I found is that he is only part right. Teachers know their classrooms best, but only through this knowledge of the classroom can someone be in a position to know what a classroom needs to make a reform successful. In other words, what Trent demonstrated is that the teachers are experts in the classroom, and thus it is best to make them experts outside the classroom to ensure successful reform.

Externalities Matter

Districts, parents and unions played a role in the success and failure of the restructuring efforts at these schools. If the union or district rules opposed a reform, the school's ability to restructure was severely impacted.

Union regulations, many times, restrict the number of hours that teachers can meet. At Web, when administrators asked teachers to stay after school to met to work on a committee, they would be reminded that the request violates the union contract. Trent increased teachers' salaried as an incentive for the extra work, but they would not have been allowed to do this without their charter status because it breaks with district policy. Moreover, if parents refused to take on some of the roles and responsibilities asked of them, the teachers and administrators could not alter their roles.

Policymakers have to put policies in place that work within the guidelines of these outside entities or that give the schools the power to work around these forces. Without this support, reforms will continue to fall short of their goals. Maybe more importantly, policymakers must consider districts, unions and parents as separate but powerful forces. If lumped together and considered as one, policymakers will, once again, lose sight of each of these entities individual impact.

New Teachers

Change is difficult for all teachers, but especially for new teachers (to the profession). Maybe schools going through restructuring should leave some of the new responsibilities optional for new teachers. Many new teachers are just figuring out how to teach and at the same time, they are being asked to lead. As a new teacher, leading is a very difficult if not impossible task.

But, in cases like Web, the small school size makes it impossible to exempt new teachers from all additional responsibilities. There just are not enough people to sit on committees if any teachers are excluded. In cases such as this, and maybe in all cases, policymakers must include enough support opportunities for new teachers to allow them to be successful. Support in the form of professional development and collegial support is important for all teachers, but, as my study demonstrates, it is especially important for teachers new to the profession. New to the profession teachers must be given professional development opportunities that focus on the school reform efforts, but also professional development in general areas such as classroom management and curriculum. Maybe more importantly, polices need to provide new teachers with opportunities for collegial help and feedback. New teachers (both) need to have colleagues available to answer questions about reforms as well as general teaching questions. Teachers need to feel free to ask questions without criticism. Humberman (1993) argues that even given the opportunity, teachers will not seek out another teacher for guidance because it would be seen as a sign of incompetence. But if teachers through policy are given a mentor that they are expected to seek out, this culture of isolation may be ended. In addition, as Rosenholtz (1991) argues, collegial feedback will reduce the uncertainties of teaching and make change possible. In short, teachers must be given the opportunity to develop teaching skills as well as to develop the skills needed to make a reform successful.

Restructuring is Demanding and Stressful

Teachers at every school described the demands put on their time, the pressure, and the stress brought about by restructuring. In short, change is difficult. At Trent, a teacher said,

> It's very demanding. It's very, very demanding. I think my biggest problem was the time that it was just, besides the classroom, you know, my job never stops. just because I get off at 2:10, I still have a ton of other things to do. And then on top of that, you still have your committee responsibilities. And then, I was going to graduate school, and I do have a life outside of that. At least, I had one before I came here. So, it is just really demanding, really time-consuming, and if your heart and mind isn't in it, then this is not the place to be because you can never ever escape your responsibilities here (Santilla, teacher, interview, Fall 1995).

Similarly at Olive Grove, teachers said things such as,

> T: What do we really have to do?" What I'm experiencing right now is some significant teacher burn-out and so are a lot of people I'm talking to. I think we're looking at a turn over at this school that hasn't been seen in ten or fifteen years. I think you're gonna see some people dropping out and I don't think it needs to be that way. But I know we all feel like we're drowning in a sea of stuff to do.

> I: The burnout is attributed to the committees or...?

> T: The burnout is attributed to everything we're trying to do--we're trying to do some of it at once.... So you take all the stuff that we're doing and you add it up and it comes out to too much. So it's a combination of committee work and other things that are contributed to the burnout (Fonsworth, teacher, interview, Spring 1997).

Policymakers must put mechanisms in place to alleviate some of the stresses of change. These mechanisms might

387

make teachers more willing to enact change. In line with Elmore's (1996) work, Trent increased teachers' salaries. This increase in salary is seen as an incentive to do more work. Incentives help to justify the long hours which may in the long run reduce stress. But, as Olive grove demonstrates, one addition such as planning time is not enough to reduce stress. Once again, policymakers need to create opportunities for multiple support mechanisms including planning time, increased salaries, support staff, and professional development that address the schools' individual needs.

Without these support mechanisms, teachers in my study had two stages of change 1) burnout and 2) movement back to the norm. Burnout was caused by the additional work without additional time or help. Teachers would begin to resent the work. After burnout, as Cuban (1993) and Lortie (1975) point out, teachers moved back to what they know. They stop anything new or innovative and revert to the teaching strategies they were familiar with. This move to constancy ends any hope of the reform in the classroom from being implemented.

Professional Development

Professional development is also important. Without proper training, reform is doomed. New and experienced teachers would have been less likely to revert to the norms of teaching with proper training.

At Web, one complaint was that all teachers were held accountable for using Integrated Thematic Instruction and Brain Compatibility, but many teachers said, "I did not receive training in ITI or Brain Compatibility until the end of the year" (teacher, focus group, Spring 1997). At Olive Grove, a teacher during a focus group complained, "We have had some pretty high powered staff development, but nothing on multiage education" (teacher, focus group, Spring 1997). Finally, at Trent, a teacher said, "We have staff development every Wednesday, [but] we have no staff development to deal with our new administrative roles" (teacher, focus group, Spring 1997).

So the main point is that teachers not only need professional development in general, but they need professional development that is linked to the goals of their school's reform. As Fullan (1991) argues, without opportunities for learning, restructuring is impossible.

Teachers Want a Say

Authors, such as Lortie (1975), have argued that teachers even given the opportunity will not increase the time spent with adults because it reduces the psychic reward that come from spending time with child. Arguments such as this seem to continue to shape policies, and thus, policymakers have not attempted to include teachers in the change process. However, my study demonstrates that in spite of the admitted difficulties and pressures, teachers want the added responsibilities that come with being a part of the main decision-making body for the school. They want to be in charge of their own destiny (Grandville, teacher, interview, Spring 1997).

Another teacher at Trent stated,

> Also, as members of committees you sit there and you know you are doing administrative...what used to be totally administrative work, you're not doing it. And it's a whole different job. You're doing the teaching, but you're also now doing the administration of.... And it's more work. You find yourself quite overworked here. But it also is a part of, or the reason why we developed what we did and what we wanted; and where we wanted to go is basically what we as a group decided is where we wanted to go and we all are a part of creating that road to it. But, it's interesting. Now we don't blame the monster out there, we blame ourselves because if something is not working it is us and it's our ?? that we have to change (teacher, focus group, Spring 1997).

Furthermore, a teacher at Trent said,

> The morale is high. To me there is such positive energy going around this school. Even though we are really bogged down with a lot of like tedious stuff. I think the ownership—the fact that teachers have been able to take ownership of the school and no longer is it the office telling us what to do. That is what I think, the high morale and the ownership that we have. And the positive light that was put in because of all the changes that have gone on (Marcos, teacher, interview, Fall 1995).

At Olive Grove, teachers repeatedly said that they "would not change the action teams or multiage education" (Rathom, teacher, interview, Fall 1995). Even at Web where there is little evidence of a change in governance, during a focus group a teacher said, "I think the good thing for me about restructuring for me is what's intellectually interesting to me is getting to talk to other people who are interested in school reform. I would like to have input" (teacher, focus group, Spring 1997).

So in opposition to the literature, and in spite of the extra work, teachers want to work together to tackle what is typically administrative work. This work gave them a feeling of ownership and control that they had not experienced previously.

388

In the end, I find that changing teachers' work is no easy task. But too often policymakers attempt to change classrooms without including the teachers or their circumstances in this change. If policymakers only take one thing from this work, I hope they remember that they cannot successfully affect the classroom without first affecting and *supporting* the teacher.

Notes

[1] Sarason (1982) and Fullan (1996) make similar arguments. Sarason's theory refers to behavior regularities and Fullan refers to second-order changes.

[2] Examples—A Nation Prepared; the subsequent development of the National Board for Professional Teaching Standards; the role played by NCTM and other professional associations in the formulation of content and performance standards, etc.

[3] The grant application process asked schools and districts to think through plans, and to rethink and create new structures and practices around six major elements of schooling which the legislation identified. From the six elements, the California Center for School Restructuring (CCSR), which was created by the California Department of Education to provide leadership, outreach and a support structure for SB 1274, assembled the regional and statewide networks of schools and districts to work on a restructuring plans that included the four goals.

[4] The data for my smaller study was taken from the three elementary "intensive" sites from the larger *School Restructuring Study*. The School Restructuring Study was a privately funded three-year investigation designed to answer the question: To what extent, and in what ways, does SB 1274 enable schools to pursue an ambitious agenda of school-wide change, with prospects for measurably affecting "powerful learning for all students?" Evidence was collected through site visits, surveys, official records, and other documents from 36 randomly selected schools. Nine of the sites —three each at the elementary, middle, and high school levels—were designated as "intensive" sites. In those schools, we made repeated visits and collected a wide range of data. In the remaining sites, we collected data through one-time site visits, school documents, and staff, parent, and student surveys.

References

Apple, M., & Jungck, S. (1992). You don't have to be a teacher to teach this unit: Teaching, technology, and control in in the classroom. In A. Hargreaves & M. Fullan (Eds.), *Understanding teacher development*. (pp. 20-42). New York: Cassell and Teacher College Press.

Archbald, D., & Porter, A. (1994). Curriculum control and teachers' perceptions of autonomy and satisfaction. *Educational evaluation and policy analysis., 16*(1), 21-39.

Barth, R. (1990). *Improving Schools from Within*. Jossey-Bass: San Francisco.

Bridges, E. (1985). *The Incompetent Teacher*. London: Falmer Press.

California Center for School Restructuring, *Primer*, 1993.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st Century*. New York: Carnegie Corporation.

Chubb, J. E. (1988). Why the current wave of school reform will fail. *The Public Interest*(90), 28-50.

Cohen, D. & Hill, H. (1997, March). *Teaching and learning mathematics in California*. Paper presented at the meeting of AERA, Chicago, IL.

Coleman, J.,& Hoffer, T. (1987). Achievement and dropout in disadvantaged and deficient families., *Public and private high school: The impact of communities*. (pp. 118-148). New York: Basic Books.

Cuban, L. (1993). *How teachers taught: Constancy and change in American Classrooms 1880-1990*. (Second Ed.). New York: Teachers College Press.

Darling-Hammond, L. (1993). Reframing the school reform agenda: Developing capacity for school transformation. *Phi Delta Kappan*(June), 753-761.

Elmore, R. (1980). Mapping backward: Using implementation analysis to structure policy decisions. *Political Science Quarterly, 44*, 601-616.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1-25.

Fullan, M. G. (1991). *The new meaning of educational change.* (Second Edition). Ontario: The Ontario Institute for Studies in Education Press (Chapters 1-6).

Gamoran, A. (1994). *Teacher empowerment: A policy in search of theory and evidence.* (Report ED 372 032). Madison: Center on Organization and Restructuring Schools.

Getzels, J. (1969). "A social psychology of education." In G. Lindzey and E. Aronson (Eds.), *The Handbook of Social Psychology* (pp. 459-471). Reading: Addison-Wesley.

Giroux, H. (1986). Educational reform and the politics of teacher empowerment. In J. Kretovics, & E. J. Nussel (Eds.), *Transforming Urban Education* (pp. 396-411). Boston: Allyn and Bacon.

Glidewell, J. C., Tucker, S., Todt, M., & Cox, S. (1983). Professional support systems: The teaching profession. In A. Nadler, J. Fisher, & B. DePaulo (Eds.), *New directions in helping* (pp. 189-212). New York: Academic Press.

Hargreaves, A. (1991). Contrived collegiality: The micropolitics of teacher collaboration. In J. Blase (Ed.), *The politics of life in schools: Power, conflict, and cooperation.* . Newbury Park: SAGE.

Hargreaves, A. (1993). Individualism and individuality: Reinterpreting the teacher culture. In J. W. Little & M. W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts.* (pp. 51-76). New York: Teachers College Press.

Huberman, M. (1993). The model of the independent artisan in teachers' professional relations. In J. W. Little & M. W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts.* (pp. 11-50). New York: Teachers College Press.

Jackson, P. (1968). *Life in classrooms.* New York: Holt, Rinehart and Winston.

Kahne, J. et al. (1991) *Restructuring: Where are we and where are we going?* Far West Lab: San Francisco ERIC ED 342 1333.

Lieberman, A. (Ed.). (1995). *The work of restructuring schools: Building from the ground up.* (Chapters 1 & 8). New York: Teachers College Press.

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record, 91*(4), 509-536.

Little, J. W. & Dorph, R. (1996) "School Restructuring Study" Interim report

Little, J. W. & Dorph, R. et al. (December,1998) *Lessons about comprehensive school reform: California's school restructuring demonstration program.* SB 1274 School Restructuring Study. University of California: Berkeley

Lortie, D. C. (1975). *Schoolteacher.* Chicago: The University of Chicago Press

Louis, K. S., Kruse, S., & Associates. (1995). *Professionalism and community: Perspectives on reforming urban school.* (Chapter 1). Thousand Oaks, California: Corwin Press, Inc.

McLaughlin, M. W. (1989). *The RAND change agent study ten years later: Macro perspectives and micro realities.* (Report No. CRC-P89-108). Center for Research on the Context of Secondary School Teaching. (ERIC Document Reproduction Service No. ED 342 085).

McNeil, L. M. (1986). *Contradictions of control: School structure and school knowledge.* New York: Routledge.

Metz, M. H. (1978). *Classroom and corridors: The crisis of authority in desegregated secondary schools.* Berkeley: University of California Press.

National Commission on Excellence in Education. (1984). *A nation at risk: The full account.* Cambridge, Massachusetts: USA Research.

Nias, J., Southworth, G., & Yeomans, R. (1989). *Staff Relationships in the primary school: A study of organizational*

cultures. (pp. 251-280). New York: Cassell.

Olsen, B., & Kirtman, L. (in press). Teacher as mediator of school reform:

An examination of teacher practice in 36 California restructuring schools. *Teachers College Record.*

Pauly, E. (1991). *The classroom crucible: What really works, what doesn't, and why.* Basic Books: New York.

Porter, A. C. (1989). External standards and good teaching: The pros and cons of telling teachers what to do. *Educational Evaluation and Policy Analysis., 11*(4), 343-356.

Rowan, B. (1990). Commitment and control: Alternative strategies for the organizational design of schools. In C. B. Cazden (Ed.), *Review of research in education.* (pp. 353-392). Washington, D. C: AERA.

Rosenholtz, S. J. (1991). *Teachers' workplace: The organization of schools.* New York: Teachers College Press.

Sarason, S. B. (1982). *The culture of school and the problem of change.* (Second Edition). New York: Allyn and Bacon.

Talbert, J. E., & McLaughlin, M. W. (1994). Teacher professionalism in local school contexts. *American Journal of Education, 102*(February), 123-153.

Tyack, D.,& Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge: Harvard University Press.

About the Author

Lisa Kirtman
Department of Elementary, Bilingual and Reading Education
California State University, Fullerton
Fullerton, CA 92834-6868
Voice: (714) 278-5901
Fax: (714) 278-3110
Email: lkirtman@fullerton.edu

Lisa Kirtman, Ph.D. is currently an Assistant Professor at California State University, Fullerton, California. Her current research interests include new teacher development and support in the area of mathematics education.